SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 1–17, January 1990
©1990 Society for Industrial and Applied Mathematics

EXISTENCE OF SOLUTIONS TO THE ELASTOHYDRODYNAMICAL EQUATIONS FOR MAGNETIC RECORDING SYSTEMS*

MICHEL CHIPOT† AND MITCHELL LUSKIN‡

Abstract. The existence of steady-state solutions to the system of nonlinear partial differential equations which are used to model the elastohydrodynamics of magnetic recording systems is demonstrated.

Key words. elastohydrodynamics, Reynolds lubrication equation, nonlinear partial differential equations

AMS(MOS) subject classifications. 35J65, 73J06

1. Introduction. The purpose of this paper is to demonstrate the existence of steady-state solutions under appropriate conditions to the system of nonlinear partial differential equations which are used to model the elastohydrodynamics of magnetic recording systems. There are two components to these mechanical systems: a medium such as a disk pressure which develops in the air bearing between the medium and the recording head causes a deflection in the medium and since the deflection of the medium influences the pressure in the air bearing.

For simplicity, we shall restrict our attention to disk systems. Let $\Omega \subset \mathbb{R}^2$ be the annular region of the disk

$$\Omega = \{ x = (x_1, x_2) \mid R < r < 1 \}$$

where $r = \sqrt{x_1^2 + x_2^2}$ and let $\Gamma \subset \Omega$ be the region where the head is in close proximity to the disk. Thus, we have scaled the spatial variables by the outer radius of the disk. The mathematical model that we use for the transverse displacement of the disk, $u = u(x,t)$, is given by [8], [13]

$$\rho\left(\frac{\partial}{\partial t} + \omega\frac{\partial}{\partial \theta}\right)^2 u = \nabla \cdot (T\nabla u) - \frac{E_p t_p^3}{12(1-\nu^2)}\Delta^2 u - \gamma\left(\frac{\partial}{\partial t} + \omega\frac{\partial}{\partial \theta}\right)u + p - p_a, \tag{1.1}$$

$$x = (x_1, x_2) \in \Omega, \qquad -\infty < t < \infty,$$

where t is time, θ is the angular coordinate in polar coordinates, ρ is area density, ω is the angular speed of rotation of the disk, T is tension, $E_p > 0$ is Young's modulus, t_p is the disk thickness, ν is Poisson's ratio, $\gamma > 0$ is the air damp coefficient, $p = p(x,t)$ is the pressure developed in the air bearing, and p_a is the ambient pressure.

It is reported in [13] that "earlier work has shown that by bonding tensioned flexible recording media to rigid support disks it is possible to have performance features

*Received by the editors December 7, 1987; accepted for publication (in revised form) March 7, 1989.

†Université de Metz, Département deMathématique et Informatique, Ile du Saulcy, 57045 Metz Cedex, France. Part of this work was done while the author was supported as a visitor at the Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota.

‡Applied Mathematics, California Institute of Technology, Pasadena, California 91125. The work of this author was supported by the National Science Foundation under grant DMS 835-1080. Part of this work was done while the author was supported as a visitor at the Institute for Mathematics and Its Applications, University of Minnesota.

similar to rigid disks, while retaining the advantages of flexible-media technology." For tensioned flexible recording media, the imposed tension is a scalar constant and the centrifugal tension is insignificant [13]. Thus, we shall take the total tension, T, to be a scalar constant. Since tensioned flexible recording media are bonded to rigid support disks at the inner and outer edge of the media (the support disk rotates with the medium), the appropriate boundary conditions are that the disk is clamped at the edges,

$$u = 0, \qquad r = R, 1, \tag{1.2}$$

and

$$\frac{\partial u}{\partial n} = \frac{\partial u}{\partial r} = 0, \qquad r = R, 1, \tag{1.3}$$

where n is the exterior normal to Ω. More details about tensioned flexible recording media and their advantages with respect to rigid (hard) disks and floppy disks are reported in [11].

The pressure, $p = p(x,t)$, is obtained from the compressible Reynolds lubrication equation [3]–[5], [8], [13]

$$\begin{aligned} &12\mu\frac{\partial(ph)}{\partial t} + 6\mu\mathbf{V}\cdot\nabla(ph) = \nabla\cdot(h^3p\nabla p), && x = (x_1,x_2)\in\Gamma, \\ &p = p_a, && x = (x_1,x_2)\in\partial\Gamma, \end{aligned} \tag{1.4}$$

where μ is the dynamic viscosity of the air, $h = h(x,t)$ is the thickness of the fluid layer between the head and the disk, and $\mathbf{V} = \mathbf{V}(x) = \omega(-x_2, x_1)$ is the velocity of the disk. We extend p to $\Omega - \Gamma$ by $p \equiv p_a$. If $\varphi = \varphi(x)$ represents the transverse coordinate of the head, then

$$h = h(u) = \varphi - u\,. \tag{1.5}$$

It will be convenient to define the dependent variables as functions of Cartesian coordinates, $x = (x_1, x_2)$, and as functions of radial coordinates, (r,θ), in different parts of this paper. However, we will denote the pressure, for example, by $p = p(x,t)$ or by $p = p(r,\theta,t)$. It will always be clear from the context which representation is appropriate.

Since $h = h(u) = \varphi - u$, the system (1.1) and (1.4) is a highly nonlinear, coupled system of partial differential equations. The physical problem requires that the variables be constrained by $p \geq 0$ and $h \geq 0$ ($h < 0$ would mean that a "head crash" has occurred). In this paper, we demonstrate the existence of a steady-state solution to (1.1)–(1.4) provided the parameters satisfy a given inequality. It is not known if there always exists a steady-state solution to the elastohydrodynamical system (1.1) and (1.4). Further, it is not known in general when unique asymptotically stable steady-state solutions exist to the elastohydrodynamical system. We note, though, that it has been demonstrated in [12] that steady-state solutions to (1.1) are asymptotically stable.

These questions of existence, uniqueness, and asymptotic stability for realistic parameter values are of practical interest in the design of magnetic recording systems. "Steady-state" solutions are often found numerically by integrating the time-dependent equations, and it can be difficult to determine whether we have converged

to a steady-state, a slowly varying transient, or a slowly varying periodic solution. In this paper, we show that steady-state solutions do exist for appropriate material constants, design parameters, and operating conditions.

In §2, we shall give some estimates for the steady-state of (1.1). In §3, we shall review the estimates that we have obtained for (1.4) in [5], and we shall give the analysis for the existence of steady-state solutions for the coupled system (1.1) and (1.4). Applications to floppy disk systems and tape systems are also given in §§2 and 3.

We suppose the reader familiar with the usual Sobolev spaces $H^1(\Omega)$, $H^k(\Omega)$, and $H_0^k(\Omega)$, and we refer to [1] for details and notation.

2. The steady-state for the deflection of the rotating disk. In this section, we shall analyze the following steady-state equation for (1.1) to find $u \in H_0^2(\Omega) \cap H^4(\Omega)$ such that

$$\rho\omega^2 \frac{\partial^2 u}{\partial \theta^2} = \nabla \cdot (T \nabla u) - E\Delta^2 u - \gamma\omega \frac{\partial u}{\partial \theta} + p, \qquad x \in \Omega, \tag{2.1}$$

where

$$E \equiv \frac{E_p t_p^3}{12(1-\nu^2)} \tag{2.2}$$

and $p \in L^2(\Omega)$.

Since $p = p(r, \theta)$ is mean square integrable, we have the Fourier expansion

$$p(r,\theta) = \sum_{m=-\infty}^{+\infty} p_m(r) e^{im\theta} \tag{2.3}$$

where the coefficients are given by

$$p_m(r) = \frac{1}{2\pi} \int_0^{2\pi} p(r,\theta) e^{-im\theta}\, d\theta. \tag{2.4}$$

It follows from the orthogonality of $e^{im\theta}$ that

$$\sum_{m=-\infty}^{+\infty} 2\pi \int_R^1 |p_m(r)|^2 r\, dr = \int_0^{2\pi} \int_R^1 |p(r,\theta)|^2 r\, dr\, d\theta < \infty$$

and, hence,

$$\int_R^1 |p_m(r)|^2 r\, dr < \infty.$$

We also assume the solution $u = u(r, \theta)$ to be mean square integrable, and we similarly expand

$$u(r,\theta) = \sum_{m=-\infty}^{+\infty} u_m(r) e^{im\theta}$$

where

$$u_m(r) = \frac{1}{2\pi} \int_0^{2\pi} u(r,\theta) e^{-im\theta}\, d\theta.$$

We again have that

$$\int_0^{2\pi} \int_R^1 |u(r,\theta)|^2 r\, dr\, d\theta = \sum_{m=-\infty}^{+\infty} 2\pi \int_R^1 |u_m(r)|^2 r\, dr < \infty.$$

Now

$$\Delta u = \left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2}\right)u,$$

so from (2.1) we obtain from matching the coefficient of $e^{im\theta}$ in both sides that

$$-\rho\omega^2 m^2 u_m = T\left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) - \frac{m^2}{r^2}\right)u_m$$

$$- E\left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) - \frac{m^2}{r^2}\right)^2 u_m - i\gamma\omega m u_m + p_m, \tag{2.5}$$

$$u_m = \frac{\partial u_m}{\partial r} = 0, \qquad r = R, 1.$$

Let $\{\varphi_{m,n}(r)\}_{n=1}^{\infty}$ be a complete set of eigenfunctions for the eigenproblem

$$-\left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) - \frac{m^2}{r^2}\right)\varphi_{m,n} + \frac{E}{T}\left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) - \frac{m^2}{r^2}\right)^2 \varphi_{m,n}$$

$$= \lambda_{m,n}\varphi_{m,n}, \qquad R < r < 1, \tag{2.6}$$

$$\varphi_{m,n} = \frac{\partial\varphi_{m,n}}{\partial r} = 0, \qquad r = R, 1,$$

which are normalized by the condition

$$\int_R^1 |\varphi_{m,n}|^2 r dr = 1$$

and where

$$0 < \lambda_{m,1} \le \lambda_{m,2} \le \cdots \le \lambda_{m,n} \le \cdots.$$

Thus, $\lambda_{m,n}$ (respectively, $\varphi_{m,n}$) are critical values (respectively, critical points) of the functional

$$J_m(\varphi) = \int_R^1 \left[\left|\frac{\partial\varphi}{\partial r}\right|^2 + \frac{m^2}{r^2}|\varphi|^2 + \frac{E}{T}\left|\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\varphi}{\partial r}\right) - \frac{m^2}{r^2}\varphi\right|^2\right] r\, dr$$

subject to the constraints that the real-valued functions, $\varphi(r)$, satisfy

$$\int_R^1 |\varphi|^2 r\, dr = 1,$$

$$\varphi(R) = \varphi(1) = 0,$$

$$\frac{\partial\varphi}{\partial r}(R) = \frac{\partial\varphi}{\partial r}(1) = 0.$$

It can be shown by classical arguments [6] that for $m = 0, \pm 1, \pm 2, \cdots$,

$$\lim_{n \to \infty} \lambda_{m,n} = +\infty. \tag{2.7}$$

We also note that $\lambda_{m,n}$ depends on R, i.e., $\lambda_{m,n} = \lambda_{m,n}(R)$ and that we can obtain from the representation of $\lambda_{m,n}$ as critical values of $J_m(\varphi)$ that $\lambda_{m,n}(R) \geq \lambda_{m,n}(0)$ for $0 \leq R < 1$. Now

$$\int_R^1 \left| \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \varphi}{\partial r} \right) - \frac{m^2}{r^2} \varphi \right|^2 r \, dr$$

$$= \int_R^1 \left[\left| \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \varphi}{\partial r} \right) \right|^2 + \frac{m^4}{r^4} |\varphi|^2 - \frac{2m^2}{r^3} \frac{\partial}{\partial r} \left(r \frac{\partial \varphi}{\partial r} \right) \varphi \right] r \, dr$$

and since φ satisfies the above boundary conditions it follows from integration by parts that

$$\int_R^1 \frac{m^2}{r^3} \frac{\partial}{\partial r} \left(r \frac{\partial \varphi}{\partial r} \right) \varphi r \, dr$$

$$= -m^2 \int_R^1 \frac{\partial \varphi}{\partial r} \frac{\partial}{\partial r} \left(\frac{\varphi}{r^2} \right) r \, dr$$

$$= -m^2 \int_R^1 r^{-2} \left| \frac{\partial \varphi}{\partial r} \right|^2 r \, dr + 2m^2 \int_R^1 r^{-3} \frac{\partial \varphi}{\partial r} \varphi r \, dr$$

$$= -m^2 \int_R^1 r^{-2} \left| \frac{\partial \varphi}{\partial r} \right|^2 r \, dr + 2m^2 \int_R^1 r^{-4} |\varphi|^2 r \, dr.$$

All the calculations in this paragraph can be used to show that

$$J_m(\varphi) = \int_R^1 \left[\left| \frac{\partial \varphi}{\partial r} \right|^2 + \frac{m^2}{r^2} |\varphi|^2 \right] r \, dr$$

$$+ \frac{E}{T} \int_R^1 \left[\left| \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial}{\partial r} \varphi \right) \right|^2 + \frac{m^4 - 4m^2}{r^4} |\varphi|^2 + \frac{2m^2}{r^2} \left| \frac{\partial \varphi}{\partial r} \right|^2 \right] r \, dr \tag{2.8}$$

for φ satisfying the imposed boundary conditions.

Classical arguments [6] can also be used to show that

$$\int_R^1 \varphi_{m,n}(r) \varphi_{m,p}(r) r \, dr = \delta_{n,p}.$$

Further, if

$$\int_R^1 |v|^2 r\,dr < \infty,$$

then the expansion

$$v(r) = \sum_{n=1}^{+\infty} v_{m,n}\varphi_{m,n}(r)$$

where

$$v_{m,n} = \int_R^1 v(r)\varphi_{m,n}(r) r\,dr$$

has the properties that

$$\int_R^1 |v|^2 r\,dr = \sum_{n=1}^{+\infty} |v_{m,n}|^2$$

and

$$\int_R^1 \left| v(r) - \sum_{n=1}^{N} v_{m,n}\varphi_{m,n}(r) \right|^2 r\,dr = \sum_{n=N+1}^{+\infty} |v_{m,n}|^2 \to 0$$

as $N \to \infty$. Finally, since

$$J_m(\varphi) \geq \frac{E}{T} \int_R^1 \frac{(m^4 - 4m^2)}{r^4} |\varphi|^2 r\,dr$$

and since

$$J_m(\varphi) \geq \int_R^1 \frac{m^2}{r^2} |\varphi|^2 r\,dr,$$

we have that

$$\lambda_{m,n} \geq \lambda_{m,1} \geq \max\left(\frac{E}{T}(m^4 - 4m^2), m^2 \right). \tag{2.9}$$

These properties of the eigenfunctions $\varphi_{m,n}(r)$ can be used to construct the expansion

$$p(r,\theta) = \sum_{m=-\infty}^{\infty} \sum_{n=1}^{\infty} p_{m,n}\varphi_{m,n}(r) e^{im\theta} \tag{2.10}$$

where

$$p_{m,n} = \frac{1}{2\pi} \int_0^{2\pi} \int_R^1 p(r,\theta)\, e^{-im\theta} \varphi_{m,n}(r) r\,dr.$$

Note that

$$\int_0^{2\pi} \int_R^1 |p|^2 r\,dr\,d\theta = 2\pi \sum_{m=-\infty}^{+\infty} \sum_{n=1}^{+\infty} |p_{m,n}|^2 < \infty.$$

We will now show that there exist unique coefficients, $u_{m,n}$, such that

$$u(r,\theta) = \sum_{m=-\infty}^{+\infty} \sum_{n=1}^{+\infty} u_{m,n}\varphi_{m,n}(r) e^{im\theta} \tag{2.11}$$

is a mean-square integrable solution to (2.1). If we formally substitute the expansions (2.10) and (2.11) into (2.1) we obtain the result that

$$-\rho\omega^2 m^2 u_{m,n} = -T\lambda_{m,n}u_{m,n} - i\gamma\omega m u_{m,n} + p_{m,n}$$

or

$$u_{m,n} = [T\lambda_{m,n} - \rho\omega^2 m^2 + i\gamma\omega m]^{-1} p_{m,n}.$$

So,

$$|u_{m,n}|^2 = [(T\lambda_{m,n} - \rho\omega^2 m^2)^2 + \gamma^2\omega^2 m^2]^{-1} |p_{m,n}|^2.$$

We note that by (2.9)

$$(T\lambda_{m,n} - \rho\omega^2 m^2)^2 + \gamma^2\omega^2 m^2 \geq \begin{cases} T^2\lambda_{0,1}^2 & \text{if } m = 0, \\ \dfrac{\gamma^2 T}{2\rho} & \text{if } m \neq 0,\ T \leq 2\rho\omega^2, \\ \dfrac{T}{2} & \text{if } m \neq 0,\ T > 2\rho\omega^2. \end{cases}$$

Thus,

$$|u_{m,n}|^2 \leq C_1 |p_{m,n}|^2 \tag{2.12}$$

where

$$C_1^{-1} = \min\left(T^2\lambda_{0,1}^2, \frac{\gamma^2 T}{2\rho}, \frac{T}{2}\right).$$

We have thus shown that if $p(r,\theta)$ is mean-square integrable, then the formal solution to (2.1)

$$u(r,\theta) = \sum_{m=-\infty}^{+\infty} \sum_{n=1}^{+\infty} u_{m,n}\varphi_{m,n}(r)e^{im\theta},$$

is unique and by (2.12) satisfies the estimate

$$\int_0^{2\pi}\int_R^1 |u(r,\theta)|^2 r\,dr\,d\theta \leq C_1 \int_0^{2\pi}\int_R^1 |p(r,\theta)|^2 r\,dr\,d\theta. \tag{2.13}$$

Since $\varphi_{m,n}$ satisfies the boundary conditions in (2.6), it follows that $u(r,\theta)$ satisfies (formally) the boundary conditions of (2.1).

Now it follows from (2.6) that $e^{im\theta}\varphi_{m,n}(r)$ are the eigenfunctions of

$$\left[-\Delta + \frac{E}{T}\,\Delta^2\right] e^{im\theta}\varphi_{m,n}(r) = \lambda_{m,n} e^{im\theta}\varphi_{m,n}(r),$$

$$e^{im\theta}\varphi_{m,n}(R) = e^{im\theta}\varphi_{m,n}(1) = 0, \tag{2.14}$$

$$\frac{\partial}{\partial r}[e^{im\theta}\varphi_{m,n}](R) = \frac{\partial}{\partial r}[e^{im\theta}\varphi_{m,n}](1) = 0.$$

Thus, since by integration by parts

$$\int_0^{2\pi}\int_R^1 [T\nabla(e^{im\theta}\varphi_{m,n})\cdot\nabla(e^{iq\theta}\varphi_{q,p}) + E\Delta(e^{im\theta}\varphi_{m,n})\Delta(e^{iq\theta}\varphi_{q,p})]r\,dr\,d\theta$$

$$= \int_0^{2\pi}\int_R^1 [(-T\Delta + E\Delta^2)(e^{im\theta}\varphi_{m,n})]e^{iq\theta}\varphi_{q,p} r\,dr\,d\theta$$

$$= T\lambda_{m,n}\int_0^{2\pi}\int_R^1 (e^{im\varphi}\varphi_{m,n})(e^{iq\theta}\varphi_{q,p})r\,dr\,d\theta$$

$$= 2\pi T\lambda_{m,n}\delta_{m,q}\delta_{n,p},$$

we also have the bound

$$\int_0^{2\pi}\int_R^1 [T|\nabla u|^2 + E|\Delta u|^2]r\,dr\,d\theta = 2\pi T\sum_{m=-\infty}^{+\infty}\sum_{n=1}^{+\infty}\lambda_{m,n}|u_{m,n}|^2$$

$$= 2\pi T\sum_{m=-\infty}^{+\infty}\sum_{n=1}^{+\infty}\frac{\lambda_{m,n}}{[(T\lambda_{m,n} - \rho\omega^2 m^2)^2 + \gamma^2\omega^2 m^2]}|p_{m,n}|^2 \tag{2.15}$$

$$\le 2\pi C_2\sum_{m=-\infty}^{+\infty}\sum_{n=1}^{+\infty}|p_{m,n}|^2 = C_2\int_0^{2\pi}\int_R^1 |p(r,\theta)|^2 r\,dr\,d\theta$$

where by (2.7) and (2.9)

$$C_2 = \max_{m,n}\left[\frac{T\lambda_{m,n}}{(T\lambda_{m,n} - \rho\omega m^2)^2 + \gamma^2 w^2 m^2}\right] < \infty.$$

Since

$$\int_0^{2\pi}\int_R^1 (-T\Delta + E\Delta^2)(e^{im\theta}\varphi_{m,n})\cdot(-T\Delta + E\Delta^2)(e^{iq\theta}\varphi_{q,p})r\,dr\,d\theta$$

$$= \int_0^{2\pi}\int_R^1 (T\lambda_{m,n}e^{im\theta}\varphi_{m,n})(T\lambda_{q,p}e^{iq\theta}\varphi_{q,p})r\,dr\,d\theta$$

$$= 2\pi T^2\lambda_{m,n}\lambda_{q,p}\delta_{m,q}\delta_{n,p},$$

we have the stronger bound

$$\int_0^{2\pi}\int_R^1 |-T\Delta u + E\Delta^2 u|^2 r\,dr\,d\theta = 2\pi T^2 \sum_{m=-\infty}^{+\infty}\sum_{n=1}^{+\infty} \lambda_{m,n}^2 |u_{m,n}|^2$$

$$= 2\pi \sum_{m=-\infty}^{+\infty}\sum_{n=1}^{+\infty} \frac{(T\lambda_{m,n})^2}{(T\lambda_{m,n} - \rho\omega^2 m^2)^2 + \gamma^2\omega^2 m^2} |p_{m,n}|^2 \tag{2.16}$$

$$\leq 2\pi C_3 \sum_{m=-\infty}^{+\infty}\sum_{n=1}^{+\infty} |p_{m,n}|^2$$

$$= C_3 \int_0^{2\pi}\int_R^1 |p(r,\theta)|^2 r\,dr\,d\theta$$

where by (2.7) and (2.9)

$$C_3 = \max\left[\frac{(T\lambda_{m,n})^2}{(T\lambda_{m,n} - \rho\omega^2 m^2)^2 + \gamma^2\omega^2 m^2}\right] < \infty.$$

The inequalities (2.13), (2.15), and (2.16) can be used to show that if $p(r,\theta)$ is mean-square integrable, then the solution $u(r,\theta)$ to (2.1) that we have constructed has the property that all of its partial derivatives of order less than or equal to four are mean-square integrable [1]. Further, this implies that all of the partial derivatives of $u(r,\theta)$ of order less than or equal to two are continuous and that the boundary conditions are satisfied in the classical sense [1].

We review the above results by the following theorem.

THEOREM 2.1. *We suppose that $E_p > 0$, $T > 0$, $\gamma > 0$, and $p \in L^2(\Omega)$. Then there exists a unique solution $u \in H_0^2(\Omega) \cap H^4(\Omega)$ to (2.1). Further, there exist positive constants C_1, C_2, and C_3 such that*

$$\int_\Omega u^2\,dx \leq C_1 \int_\Omega p^2\,dx, \tag{2.17}$$

$$\int_\Omega [T|\nabla u|^2 + E|\Delta u|^2]\,dx \leq C_2 \int_\Omega p^2\,dx, \tag{2.18}$$

$$\int_\Omega |-T\Delta u + E\Delta^2 u|^2\,dx \leq C_3 \int_\Omega p^2\,dx. \tag{2.19}$$

The constants C_1, C_2, and C_3 can be chosen independent of R. Also, the constant C_1, can be chosen independent of ω.

A more detailed analysis of $\lambda_{m,n}$ and the constant C_2 can be used to demonstrate that C_2 is independent of ω for $\rho\omega^2 \leq T$. However, we prefer to give the following elementary proof.

PROPOSITION 2.2. *Assume that*

$$\rho\omega^2 \leq T. \tag{2.20}$$

If $p \in L^2(\Omega)$, then there exists a unique solution to

$$\begin{aligned} &E\Delta^2 u - T\Delta u + \rho\omega^2 \frac{\partial^2 u}{\partial \theta^2} + \gamma\omega \frac{\partial u}{\partial \theta} = p \quad \text{in } \Omega, \\ &u \in H_0^2(\Omega). \end{aligned} \tag{2.21}$$

Moreover, there exists a constant $C_4 = C_4(E)$, independent of ω, such that

$$|u|_{H_0^2(\Omega)} \leq C_4 |p|_{L^2(\Omega)}. \tag{2.22}$$

Proof. This is a straightforward application of the Lax–Milgram theorem. Indeed, consider the weak formulation of (2.21), i.e., set

$$a(u,w) = \int_\Omega E\Delta u \cdot \Delta w + T\nabla u \cdot \nabla w - \rho\omega^2 \frac{\partial u}{\partial \theta} \cdot \frac{\partial w}{\partial \theta} + \gamma\omega \frac{\partial u}{\partial \theta} \cdot w \, dx.$$

Then clearly $a(u,w)$ is a bilinear, continuous form on $H_0^2(\Omega)$. Moreover,

$$\begin{aligned} a(u,u) &= \int_\Omega E\Delta u \cdot \Delta u + T\nabla u \cdot \nabla u - \rho\omega^2 \left(\frac{\partial u}{\partial \theta}\right)^2 + \gamma \frac{w}{2} \frac{\partial}{\partial \theta} u^2 \, dx \\ &= \int_\Omega E\Delta u \cdot \Delta u + T\nabla u \cdot \nabla u - \rho\omega^2 \left(\frac{\partial u}{\partial \theta}\right)^2 dx. \end{aligned} \tag{2.23}$$

(We used the fact that since $\Omega \subset \mathbb{R}^2$, $H_0^2(\Omega) \subset C(\bar{\Omega})$ and $u(r,\theta) = u(r,\theta + 2\pi)$.) Now, recall that

$$\begin{aligned} \frac{\partial u}{\partial \theta} &= \frac{\partial u}{\partial x_1} \cdot \frac{\partial x_1}{\partial \theta} + \frac{\partial u}{\partial x_2} \cdot \frac{\partial x_2}{\partial \theta} \\ &= -r \sin\theta \frac{\partial u}{\partial x_1} + r \cos\theta \frac{\partial u}{\partial x_2}. \end{aligned}$$

Hence, by the Cauchy–Schwarz inequality,

$$\begin{aligned} \left(\frac{\partial u}{\partial \theta}\right)^2 &= r^2 \left(-\sin\theta \frac{\partial u}{\partial x_1} + \cos\theta \frac{\partial u}{\partial x_2}\right)^2 \\ &\leq |\nabla u|^2 \qquad \text{a.e. on } \Omega. \end{aligned}$$

Recalling (2.23) we obtain

$$\begin{aligned} a(u,u) &= \int_\Omega E\Delta u \cdot \Delta u + (T - \rho\omega^2)|\nabla u|^2 \, dx \\ &\geq E \int_\Omega (\Delta u)^2 \, dx. \end{aligned}$$

Since E is assumed to be positive, and since

$$\int_\Omega (\Delta u)^2 \, dx$$

defined on $H_0^2(\Omega)$ is a norm equivalent to the usual one, $a(u, w)$ is a bilinear, continuous, coercive form on $H_0^2(\Omega)$. Now, for $p \in L^2(\Omega)$ it is clear that

$$w \longmapsto \int_\Omega pw \, dx$$

is a continuous linear form on $H_0^2(\Omega)$, so, by the Lax–Milgram theorem there is a unique u in $H_0^2(\Omega)$ satisfying

$$a(u, w) = \int_\Omega pw \, dx \quad \forall w \in H_0^2(\Omega).$$

Moreover, taking $w = u$ in the above equality, we can easily see that (2.22) holds.

It is easy to see that u satisfies (2.21) in the distributional sense. So, we have

$$\Delta^2 u \in L^2(\Omega)$$

since all the other functions appearing in (2.21) are in $L^2(\Omega)$. Hence, by well-known results $u \in H^4(\Omega)$. In particular we recover the fact that the condition

$$u = \frac{\partial u}{\partial r} = 0 \qquad \text{on } \partial\Omega$$

holds in the usual sense (see [1], [7] for details). □

We could have assumed p in the dual of $H_0^2(\Omega)$ and the existence of a weak solution to (2.21) would still have held true.

We note that (2.20) is the condition that (2.21) be elliptic when $E \equiv 0$. In the case that $\gamma \equiv 0$, existence and uniqueness can fail when $\rho\omega^2 > T$. If $\gamma \equiv 0$, then

$$(T\lambda_{m,n} - \rho\omega^2 m^2) u_{m,n} = p_{m,n}.$$

Hence, if $T\lambda_{m,n} = \rho\omega^2 m^2$ for some m, n, then there exist pressures, p, such that (2.21) does not have a solution. One example is clearly

$$p(r, \theta) = e^{im\theta} \varphi_{m,n}(r).$$

Also, in this case ($T\lambda_{m,n} = \rho\omega^2 m^2$ for some m, n) solutions to (2.21) are not unique since if $u(r, \theta)$ is a solution to (2.21), then

$$u(r, \theta) + e^{im\theta} \varphi_{m,n}(r)$$

is also a solution to (2.21). However, Theorem 2.1 guarantees existence and uniqueness when $\gamma > 0$ for all ω.

We note that for the floppy disk, the outer edge is not bonded to a rigid support disk. In this case, the tension depends on r. Furthermore, the radial tension coefficient vanishes at the outer edge [2]. Although clamped plate boundary conditions are appropriate at the inner edge, the plate is free at the outer edge [2]. The head in a floppy disk system does not fly above the medium on an air bearing. In this case, p in (1.1) represents the load on the disk from contact with the head. The analysis given for Proposition 2.2 applies to the floppy disk system if E is sufficiently large. In this case it is no longer true that $\rho\omega^2 \leq T$ everywhere in Ω. However, for $E/\rho\omega^2$ sufficiently large, we can use the inequality

$$\rho\omega^2 \int_\Omega \left(\frac{\partial u}{\partial \theta} \right)^2 dx \leq \frac{E}{2} \int_\Omega (\Delta u)^2 \, dx$$

for $u \in H^2(\Omega)$ and $u = \dfrac{\partial u}{\partial n} = 0$ on the inner edge of the disk.

3. The coupled problem. The steady-state solution of (1.1), (1.4) is

$$E\Delta^2 u - T\Delta u + \rho\omega^2 \frac{\partial^2 u}{\partial \theta^2} + \gamma\omega \frac{\partial u}{\partial \theta} = p - p_a, \qquad \text{in } \Omega, \tag{3.1}$$

$$u \in H_0^2(\Omega),$$

$$\nabla \cdot (h(u)^3 p \nabla p) = 6\mu \mathbf{V} \cdot \nabla(ph(u)), \qquad \text{in}\Gamma, \tag{3.2}$$

$$p = p_a \qquad \text{on } \partial\Gamma,$$

where $\mathbf{V} = \omega(-x_2, x_1)$, $\Omega = \{ x = (x_1, x_2) \mid R < r < 1 \}$.

We shall restrict p so that $p \geq 0$ and set $v = p^2$. The problem then becomes to find (u, v) such that

$$E\Delta^2 u - T\Delta u + \rho\omega^2 \frac{\partial^2 u}{\partial \theta^2} + \gamma\omega \frac{\partial u}{\partial \theta} = \sqrt{v} - \sqrt{v_a}, \qquad \text{in } \Omega, \tag{3.3}$$

$$u \in H_0^2(\Omega),$$

$$\nabla \cdot (h(u)^3 \nabla v) = \omega \mathbf{W} \cdot \nabla(\sqrt{v} \quad h(u)), \qquad \text{in } \Gamma, \tag{3.4}$$

$$v = v_a, \quad \text{on } \partial\Gamma,$$

where we have set $v_a = p_a^2$, $\mathbf{W} = 12\mu(-x_2, x_1)$.

We are going to solve (3.4) in a weak sense. Noting that $\nabla \cdot \mathbf{W} = 0$ we see that if v satisfies (3.4)—say in a classical sense—we have

$$\int_\Omega h(u)^3 \nabla v \cdot \nabla \xi \, dx - \int_\Omega \omega\sqrt{v}\, h(u) \mathbf{W} \cdot \nabla \xi \, dx = 0 \quad \forall \xi \in H_0^1(\Omega), \tag{3.5}$$

$$v = v_a \quad \text{on } \partial\Omega.$$

From the Sobolev embedding theorem [1], [7] we have that

$$H_0^2(\Omega) \subset C(\bar{\Omega})$$

with continuous inclusion. So, for some positive constant C_5 we have that

$$|w|_{L^\infty(\Omega)} \leq C_5 |w|_{H_0^2(\Omega)}, \qquad w \in H_0^2(\Omega).$$

Let us now collect our assumptions. We assume that the function φ (see (1.5)) satisfies

$$\begin{aligned} &\varphi \text{ is a Lipschitz continuous function on } \Gamma, \\ &0 < m \leq \varphi(x) \leq M \quad \text{a.e. for } x \in \Gamma, \end{aligned} \tag{3.6}$$

where m and M are two positive constants.

Assume also that

$$\Gamma \subset \Omega \text{ is a domain of } \mathbf{R}^2 \text{ with Lipschitz boundary;} \tag{3.7}$$

let $|\Gamma|$ denote the Lebesgue measure of Γ and let d denote the smallest width of a strip containing Γ. Then we can prove the following theorem.

THEOREM 3.1. *Let m' be any positive number such that*

$$0 < m' < m.$$

There exists a solution (u, v) of (3.3), (3.4) *if* (3.6) *and* (3.7) *hold and if*

$$\left(\frac{a + \sqrt{a^2 + 4b}}{2}\right) \leq p_a \frac{(m - m')}{C_4 C_5} \tag{3.8}$$

where

$$a = \frac{[12 d\omega\mu(M + m - m')]^2 |\Gamma|^{1/2}}{(m')^6}, \qquad b = a|\Gamma|^{1/2} p_a^2.$$

Proof. Set

$$K_{\mathscr{R}} = \{ v \in L^2(\Gamma) \mid v \geq 0 \text{ a.e. on } \Gamma,\ |v - v_a|_{L^2(\Gamma)} \leq \mathscr{R} \} \tag{3.9}$$

where $\mathscr{R}$ is a positive real number that we will choose later on. It is clear that $K_{\mathscr{R}}$ is a closed convex set of $L^2(\Gamma)$.

For $v \in K_{\mathscr{R}}$ we have

$$\sqrt{v} - \sqrt{v_a} \in L^2(\Omega) \tag{3.10}$$

(v is, of course, supposed to be extended by v_a outside of Γ). Indeed, this is an easy consequence of the inequality

$$|\sqrt{v} - \sqrt{v_a}| \leq \frac{1}{\sqrt{v_a}} |v - v_a|. \tag{3.11}$$

So, by Theorem 2.1 and Proposition 2.2 there exists a unique solution u of

$$\begin{aligned} &E\Delta^2 u - T\Delta u + \rho\omega^2 \frac{\partial^2 u}{\partial\theta^2} + \gamma\omega \frac{\partial u}{\partial\theta} = \sqrt{v} - \sqrt{v_a}, \qquad \text{in } \Omega, \\ &u \in H_0^2(\Omega), \end{aligned} \tag{3.12}$$

and there is a constant $C_4 = C_4(E)$ such that

$$|u|_{H_0^2(\Omega)} \leq \frac{C_4}{\sqrt{v_a}} |v - v_a|_{L^2(\Gamma)} \tag{3.13}$$

(we use here the fact that, by (3.11), we have $|\sqrt{v} - \sqrt{v_a}|_{L^2(\Gamma)} \leq |v - v_a|_{L^2(\Gamma)}/\sqrt{v_a}$). The constant C_4 is independent of ω for $\rho\omega^2 \leq T$. Let m' be any positive number such that

$$0 < m' < m.$$

Let us show first that we can select $\mathscr{R}$ in such a way that

$$h(u) = \varphi - u \geq m' > 0. \tag{3.14}$$

Recall from the Sobolev embedding theorem given above that

$$|u|_{L^\infty(\Omega)} \le C_5 |u|_{H_0^2(\Omega)}$$

$$\le \frac{C_4 C_5}{\sqrt{v_a}} |v - v_a|_{L^2(\Gamma)} \qquad \text{(by (3.13))}.$$

Now, by (3.6), (3.14) will hold if

$$|u|_\infty \le m - m' \tag{3.15}$$

So (3.14) will hold if

$$\frac{C_4 C_5 \mathscr{R}}{\sqrt{v_a}} \le m - m'$$

or, equivalently,

$$\mathscr{R} \le \sqrt{v_a} \frac{(m - m')}{C_4 C_5}. \tag{3.16}$$

Assume that $\mathscr{R}$ has been chosen such that (3.16) holds. Then since $h(u)$ is strictly positive there exists a unique function $\mathscr{T}(v) \ge 0$ which is a solution of

$$\int_\Gamma h(u)^3 \nabla \mathscr{T}(v) \cdot \nabla \xi \, dx - \omega \int_\Gamma \sqrt{\mathscr{T}(v)} h(u) \mathbf{W} \cdot \nabla \xi \, dx = 0, \qquad \forall \xi \in H_0^1(\Omega),$$

$$\mathscr{T}(v) = v_a, \quad \text{on } \Gamma \tag{3.17}$$

(we refer to [5] for a proof of this result).

If we can prove that $\mathscr{T}(v)$ has a fixed point we will be done. First, let us prove that for a suitable choice of $\mathscr{R}$, T maps $K_{\mathscr{R}}$ into itself. Indeed, we already know that $\mathscr{T}(v) \ge 0$. Next, if we take $\xi = \mathscr{T}(v) - v_a$ in (3.17) we get

$$(m')^3 |\nabla(\mathscr{T}(v) - v_a)|^2_{L^2(\Gamma)} \le \int_\Gamma h^3(u) |\nabla(\mathscr{T}(v) - v_a)|^2 \, dx$$

$$= \omega \int_\Gamma h(u) \sqrt{\mathscr{T}(v)} \mathbf{W} \cdot \nabla(\mathscr{T}(v) - v_a) \, dx.$$

Using the Cauchy–Schwarz inequality and recalling (3.6), (3.14) we deduce

$$(m')^3 |\nabla(\mathscr{T}(v) - v_a)|^2_{L^2(\Gamma)} \le 12\omega\mu(M + m - m') |\nabla(\mathscr{T}(v) - v_a)|_{L^2(\Gamma)} \left(\int_\Gamma \mathscr{T}(v) \, dx \right)^{1/2}$$

(note that $\sup_\Gamma |\mathbf{W}| \le 12\mu$). Hence, we have that

$$(m')^3 |\nabla(\mathscr{T}(v) - v_a)|_{L^2(\Gamma)} \le 12\omega\mu(M + m - m') \left(\int_\Gamma \mathscr{T}(v) \, dx \right)^{1/2}$$

$$\le 12\omega\mu(M + m - m') \left\{ \int_\Gamma |\mathscr{T}(v) - v_a| \, dx + |\Gamma| v_a \right\}^{1/2}. \tag{3.18}$$

By the Poincaré inequality,

$$(m')^3|\mathcal{T}(v) - v_a|_{L^2(\Gamma)} \leq d(m')^3|\nabla(\mathcal{T}(v) - v_a)|_{L^2(\Gamma)}$$

$$\leq 12d\omega\mu(M + m - m')\{|\Gamma|^{1/2}|\mathcal{T}(v) - v_a|_{L^2(\Gamma)} + |\Gamma|v_a\}^{1/2} \tag{3.19}$$

where d denotes the smallest width of a strip containing Γ. So, we get

$$|\mathcal{T}(v) - v_a|^2_{L^2(\Gamma)} \leq a|\mathcal{T}(v) - v_a|_{L^2(\Gamma)} + b$$

with

$$a = \frac{[12d\omega\mu(M + m - m')]^2|\Gamma|^{1/2}}{(m')^6}, \qquad b = \frac{[12d\omega\mu(M + m - m')]^2|\Gamma|v_a}{(m')^6} \tag{3.20}$$

and $\mathcal{T}$ maps $K_{\mathcal{R}}$ into itself provided that

$$\left(\frac{a + \sqrt{a^2 + 4b}}{2}\right) \leq \mathcal{R}. \tag{3.21}$$

Assume that (see (3.16))

$$\left(\frac{a + \sqrt{a^2 + 4b}}{2}\right) < \sqrt{v_a}\frac{(m - m')}{C_4C_5} \tag{3.22}$$

then we can select $\mathcal{R}$ such that (3.16) and (3.21) hold. Thus for $v \in K_{\mathcal{R}}$, $\mathcal{T}(v) \in K_{\mathcal{R}}$. Now, from (3.18) it is clear that $\mathcal{T}(K_{\mathcal{R}})$ is relatively compact in $K_{\mathcal{R}}$ (since $H^1(\Omega)$ is compactly embedded in $L^2(\Omega)$, see [7]). So, provided we prove that $\mathcal{T}$ is continuous on $K_{\mathcal{R}}$, by the Schauder fixed point theorem (see [7]) we can conclude the existence of (u, v) satisfying (3.3), (3.5).

To prove the continuity of $\mathcal{T}$ we proceed as follows: let $v_n \in K_{\mathcal{R}}$ be such that $v_n \to v$ in $L^2(\Gamma)$. Let us denote by u_n the solution of (3.3) corresponding to $v = v_n$, and by u the one corresponding to v. From Theorem 2.1 and Proposition 2.2 we derive easily

$$|u_n - u|_{H^2_0(\Omega)} \leq C|(\sqrt{v_n} - \sqrt{v_a}) - (\sqrt{v} - \sqrt{v_a})|_{L^2(\Gamma)}$$

$$\leq C|\sqrt{v_n} - \sqrt{v}|_{L^2(\Gamma)} \leq C\left|\sqrt{|v_n - v|}\right|_{L^2(\Gamma)}$$

$$= C\left(\int_\Gamma |v_n - v|\,dx\right)^{1/2} \leq C|\Gamma|^{1/4}|v_n - v|^{1/2}_{L^2(\Gamma)}.$$

(We used the Cauchy–Schwarz inequality.) Hence $u_n \to u$ in $H^2_0(\Omega)$ and also uniformly on $\bar{\Omega}$. (Recall that $H^2_0(\Omega) \subset C(\bar{\Omega})$ continuously.)

Now, from (3.19) we deduce that for some constant C independent of n we have

$$|\mathcal{T}(v_n)|_{H^1(\Gamma)} \leq C.$$

So we can extract a subsequence n_k from n such that

$$\mathscr{T}(v_{n_k}) \to w \quad \text{in } H^1(\Gamma), \qquad \mathscr{T}(v_{n_k}) \to w \quad \text{in } L^2(\Gamma). \tag{3.23}$$

If we let k go to $+\infty$ in the equality

$$\int_\Gamma h^3(u_{n_k})\nabla\mathscr{T}(v_{n_k})\cdot\nabla\xi - \omega\sqrt{\mathscr{T}(v_{n_k})}h(u_{n_k})\mathbf{W}\cdot\nabla\xi\,dx = 0 \qquad \forall\xi\in H_0^1(\Omega),$$

we obtain (recall that $h^3(u_{n_k}) \to h^3(u)$ uniformly)

$$\begin{aligned} &\int_\Gamma h^3(u)\nabla w\cdot\nabla\xi - \omega\sqrt{w}h(u)\mathbf{W}\cdot\nabla\xi\,dx = 0 \qquad \forall\xi\in H_0^1(\Omega),\\ &w = v_a \qquad \text{on } \partial\Gamma. \end{aligned} \tag{3.24}$$

By uniqueness of the solution of such a problem we have $w = \mathscr{T}(v)$ (see (3.17)). It results from (3.23) that the whole sequence $\mathscr{T}(v_n)$ converges toward $\mathscr{T}(v)$. This proves that $\mathscr{T}(v_n)$ converges toward $\mathscr{T}(v)$. Hence, $\mathscr{T}$ is continuous on $K_{\mathscr{R}}$ and this completes the proof of the theorem.

COROLLARY 3.1. *If ω or $|\Gamma|$ are small enough or m is large enough with M/m fixed and* (3.6) *and* (3.7) *hold, then there exists a solution (u,v) of* (3.3), (3.4).

Proof. Clearly, (3.8) holds for ω or $|\Gamma|$ small enough, all other quantities being kept fixed. Also, (3.8) holds if we set $m' = m/2$ and m is large enough with M/m fixed. □

It should be possible to apply the techniques of this paper to prove results similar to Theorem 3.1 for tape systems. Tape systems are usually modeled by a simplified shell model for the displacement of the tape and the compressible Reynolds lubrication equation for the air bearing [9].

REFERENCES

[1] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, New York, 1965.

[2] R. BENSON AND D. BOGY, *Deflection of a very flexible spinning disk due to a stationary transverse load*, ASME J. Appl. Mech., 45 (1978), pp. 636–642.

[3] M. CHIPOT, *On the Reynolds lubrication equation*, Nonlinear Anal., 12 (1988), pp. 699–718.

[4] M. CHIPOT AND M. LUSKIN, *The compressible Reynolds lubrication equation*, Metastability and Incompletely Posed Problems, IMA Volumes in Mathematics and its Applications 3, S. Antman, J. Ericksen, D. Kinderlehrer, and I. Muller, eds., Springer-Verlag, Berlin, New York, 1987, pp. 61–76.

[5] ———, *Existence and uniqueness of solutions to the compressible Reynolds lubrication equation*, SIAM J. Math. Anal., 17 (1986), pp. 1390–1399.

[6] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics* 1, Interscience, New York, 1952.

[7] D. GILBARG AND N.S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Second edition, Springer-Verlag, Berlin, New York, 1985.

[8] H.J. GREENBERG, *Flexible disk-read/write head interface*, IEEE Trans. Magn., MAG–14 (1978), p. 336.

[9] ———, *Study of head-tape interaction in high speed rotating head recording*, IBM J. Res. Develop., 23 (1979), pp. 197–205.

[10] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities*, Academic Press, New York, 1980.

[11] J. KNUDSEN, *Stretched-surface recording disk for use with a flying head*, IEEE Trans. Magn., MAG–21 (1985), pp. 2588–2591.

[12] M. LUSKIN, *A theoretical analysis of a mathematical model for the deflection of a spinning disk*, University of Minnesota Mathematics Report 84-144, 1984.

[13] D. PERRY, J. KNUDSEN, AND W. SKELCHER, *Design of flying heads for tensioned flexible recording media*, IEEE Trans. Magn., MAG–22 (1986), pp. 1005–1007.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 18–36, January 1990
© 1990 Society for Industrial and Applied Mathematics
002

A QUASI-VARIATIONAL INEQUALITY ARISING IN ELASTOHYDRODYNAMICS*

BEI HU†

Abstract. In this paper, a quasi-variational inequality arising in elastohydrodynamic lubrication is studied. In the two-dimensional case modeling a thin fluid film between an elastic ball and a plane, the existence of a smooth solution is proved provided that the viscosity is assumed to be constant. In this case, estimates for the support of the solution are also established and uniqueness of the solution is also proved under some restrictions. In the case where the viscosity is not constant, the existence, regularity, and uniqueness are proved under additional restrictions. Finally, for the one-dimensional problem describing a thin fluid between a rolling cylinder and a plane, the fact that the free boundary consists of at most one point is established in addition to existence and uniqueness.

Key words. variational inequality, free boundary problem, a priori estimates, fixed point

AMS(MOS) subject classifications. 35R35, 35B45, 35B65

1. The model. The lubrication of a ball rolling in the positive x direction gives rise to a variational inequality:

$$-\frac{\partial}{\partial x}\left(\frac{\rho\tilde{h}^3}{12\mu}\frac{\partial\tilde{u}}{\partial x}\right)-\frac{\partial}{\partial y}\left(\frac{\rho\tilde{h}^3}{12\mu}\frac{\partial\tilde{u}}{\partial y}\right)\geq-\frac{\partial}{\partial x}\left(\frac{\rho v\tilde{h}}{2}\right) \tag{1.1}$$

$$\tilde{u}\geq 0 \tag{1.2}$$

$$\tilde{u}\cdot\left[-\frac{\partial}{\partial x}\left(\frac{\rho\tilde{h}^3}{12\mu}\frac{\partial\tilde{u}}{\partial x}\right)-\frac{\partial}{\partial y}\left(\frac{\rho\tilde{h}^3}{12\mu}\frac{\partial\tilde{u}}{\partial y}\right)+\frac{\partial}{\partial x}\left(\frac{\rho v\tilde{h}}{2}\right)\right]=0 \tag{1.3}$$

where $\tilde{u}$ is the pressure, v is the average surface speed ($v>0$), ρ is the density of the liquid which shall be assumed to be constant, $\mu=\mu(\tilde{u})$ is the viscosity coefficient of the liquid, and $\tilde{h}$ is the film thickness which takes the form:

$$\tilde{h}(x,y)=\tilde{k}+\frac{x^2+y^2}{2R}+\frac{2}{\pi E'}\int\frac{\tilde{u}(s,t)dsdt}{\sqrt{(x-s)^2+(y-t)^2}} \tag{1.4}$$

where E' is the effective modulus, and $\tilde{k}$ is a positive constant.

The variational inequality (1.1)–(1.3) (with $\tilde{h}$ a given function) occurs in a simplified model of a lubrication problem (see [2]); the dependence of $\tilde{h}$ on the pressure, as in (1.4), assumes that the ball is elastic; this is the case when the load is large. The system (1.1)–(1.4) forms an elastohydrodynamics lubrication model; for more details see [3], [6]. In this paper, we study the quasi-variational inequality (1.1)–(1.4) in a bounded, but large domain Ω. Since $\tilde{u}$ is small on $\partial\Omega$, it seems natural to impose the boundary condition

$$\tilde{u}=0 \quad \text{on } \partial\Omega. \tag{1.5}$$

If μ is constant, then setting the new variables

$$h=2R\tilde{h},\quad u=\frac{4R}{\pi E'}\tilde{u},\quad k=2R\tilde{k},\quad \lambda=\frac{12\mu v(2R)^3}{\pi E'} \tag{1.6}$$

* Received by the editors August 15, 1988; accepted for publication (in revised form) March 21, 1989.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

we can rewrite (1.1)–(1.5) in the form

$$-\nabla(h^3\nabla u) \geq -\lambda\frac{\partial h}{\partial x} \quad \text{for } (x,y)\in\Omega \tag{1.7}$$

$$u \geq 0 \quad \text{for } (x,y)\in\Omega \tag{1.8}$$

$$u\cdot\left[-\nabla(h^3\nabla u)+\lambda\frac{\partial h}{\partial x}\right]=0 \quad \text{for } (x,y)\in\Omega \tag{1.9}$$

$$u=0 \quad \text{for } (x,y)\in\partial\Omega \tag{1.10}$$

and

$$h(x,y)=k+x^2+y^2+\int_\Omega \frac{u(s,t)dsdt}{\sqrt{(x-s)^2+(y-t)^2}}. \tag{1.11}$$

In §3 we prove the existence of a $C^{1,1}$ solution by using the fixed point theorem; the proof uses some estimates derived in §2. In §4 we take Ω to be a disc with large radius M and obtain some estimates for the support of the solution, i.e., we prove that for some small $\epsilon>0$,

$$u(x,y)>0 \qquad \text{for } -M<x<-\epsilon M. \tag{1.12}$$

In §5 we prove the uniqueness of the solution provided λ is small. In §6 we study the problem (1.7)–(1.11) in the case where

$$\mu=\mu_0 e^{\alpha u}, \qquad \mu_0>0,\ \alpha>0; \tag{1.13}$$

this case is of particular physical interest (see [1], [6]). In [7], by a penalty formulation of the quasi-variational inequality, the existence of $H_0^1(\Omega)$ solution has been proved provided α is small. On the other hand, we are going to prove the existence of a $C^{1,1}$ solution provided α is small (the case $\alpha=0$ is treated in §3).

In §7 we consider a rolling cylinder instead of a rolling ball; the lubrication problem then reduces to a one-dimensional quasi-variational inequality:

$$-\left(\tilde{h}^3\frac{\tilde{u}'}{\mu}\right)' \geq -6v\tilde{h}' \quad \text{for } x\in[-M,M] \tag{1.14}$$

$$\tilde{u}\geq 0 \quad \text{for } x\in[-M,M] \tag{1.15}$$

$$\tilde{u}\cdot\left[-\left(\tilde{h}^3\frac{\tilde{u}'}{\mu}\right)'+6v\tilde{h}'\right]=0 \quad \text{for } x\in[-M,M] \tag{1.16}$$

$$\tilde{u}(\pm M)=0 \tag{1.17}$$

and

$$\tilde{h}=\tilde{k}+\frac{x^2}{2R}+\frac{4}{\pi E'}\int_{-M}^{M}\tilde{u}(s)\log\frac{2M}{|x-s|}ds. \tag{1.18}$$

Setting analogously to (1.6)

$$h=2R\tilde{h}, \quad u=\frac{8R}{\pi E'}\tilde{u}, \quad k=2R\tilde{k}, \quad \lambda=\frac{24\mu_0 v(2R)^3}{\pi E'}, \tag{1.19}$$

the problem, with μ given as in (1.13), reduces to

$$-\left(h^3\frac{u'}{e^{\alpha u}}\right)' \geq -\lambda h' \quad \text{for } x \in [-M, M] \tag{1.20}$$

$$u \geq 0 \quad \text{for } x \in [-M, M] \tag{1.21}$$

$$u \cdot \left[-\left(h^3\frac{u'}{e^{\alpha u}}\right)' + \lambda h'\right] = 0 \quad \text{for } x \in [-M, M] \tag{1.22}$$

$$u(\pm M) = 0 \tag{1.23}$$

and

$$h = k + x^2 + \int_{-M}^{M} u(s) \log \frac{2M}{|x - s|} ds. \tag{1.24}$$

Assuming α to be small, we prove the existence of a $C^{1,1}$ solution. It is also proved that the solution is unique and that the free boundary consists of at most one point provided λ is small.

Numerical work for this case can be found in [1]. In [7], numerical work has been done based on a penalty formulation of the variational inequality.

2. A priori estimates. Later on we shall need some estimates for the solution of the quasi-variational inequality

$$-\nabla(h^3\nabla u) \geq -\lambda N\frac{\partial h}{\partial x} \quad \text{for } (x, y) \in \Omega \tag{2.1}$$

$$u \geq 0 \quad \text{for } (x, y) \in \Omega \tag{2.2}$$

$$u \cdot \left[-\nabla(h^3\nabla u) + \lambda N\frac{\partial h}{\partial x}\right] = 0 \quad \text{for } (x, y) \in \Omega \tag{2.3}$$

$$u = 0 \quad \text{for } (x, y) \in \partial\Omega \tag{2.4}$$

and

$$h = k + N^2(x^2 + y^2) + N\int_{\Omega} \frac{u(s, t)dsdt}{\sqrt{(x - s)^2 + (y - t)^2}} \tag{2.5}$$

where λ, k, N are positive constants, $\lambda \leq \Lambda$, and Ω is a smooth domain in R^2.

Remark. Later on we shall take N to be different constants in the proof of the existence and the estimation of the support of the solution.

LEMMA 2.1. *Assume that* (u, h) *is a solution of* (2.1)-(2.5) *with* $u \in W_0^{1,2}(\Omega)$, $h \in W^{1,2}(\Omega)$. *Then*

$$\int_{\Omega} h^3|\nabla u|^2 dxdy \leq \frac{\lambda^2|\Omega|}{k}N^2. \tag{2.6}$$

Proof. Integrating (2.3) over Ω, we get

$$
\begin{aligned}
\int_\Omega h^3|\nabla u|^2 dxdy &= -\lambda N \int_\Omega u\frac{\partial h}{\partial x} \\
&= \lambda N \int_\Omega h\frac{\partial u}{\partial x} \\
&\le \lambda N \left(\int_\Omega h^2 \left|\frac{\partial u}{\partial x}\right|^2 dxdy\right)^{1/2} \left(\int_\Omega dxdy\right)^{1/2} \\
&\le \frac{\lambda N|\Omega|^{1/2}}{k^{1/2}} \left(\int_\Omega h^3|\nabla u|^2 dxdy\right)^{1/2}
\end{aligned} \tag{2.7}
$$

and hence (2.6) follows. □

Extend u by zero outside Ω. Then $u \in W_0^{1,2}(R^2)$ and, by a change of variables,

$$
h(x,y) = k + N^2(x^2+y^2) + N\int_{R^2} \frac{1}{\sqrt{s^2+t^2}} u(x-s, y-s)dsdt. \tag{2.8}
$$

Thus

$$
\nabla h(x,y) = (2N^2x, 2N^2y) + N\int_{R^2} \frac{1}{\sqrt{s^2+t^2}} \nabla u(x-s, y-t)dsdt. \tag{2.9}
$$

LEMMA 2.2. *Under the assumption of Lemma* 2.1, *we have*

$$
\|\nabla h\|_{L^p(\Omega)} \le CN^2 \tag{2.10}
$$

where $2 < p < \infty$, *and* C *is a constant depending on* Ω , p, *and* Λ.

Proof. Applying Young's inequality (see, for example, [5, Lemma 7.12]) to (2.9), we get

$$
\|\nabla h\|_{L^p(\Omega)} \le C(\Omega)N^2 + C(\Omega, p)N\|\nabla u\|_{L^2(\Omega)}. \tag{2.11}
$$

Since, by (2.6),

$$
k^3 \int_\Omega |\nabla u|^2 dxdy \le \frac{\lambda^2|\Omega|}{k} N^2 \tag{2.12}
$$

and therefore,

$$
\|\nabla u\|_{L^2(\Omega)} \le \left(\frac{\lambda^2|\Omega|}{k^4} N^2\right)^{1/2} = \frac{\lambda|\Omega|^{1/2}}{k^2} N. \tag{2.13}
$$

Substituting this into (2.11), (2.10) follows. □

Setting

$$
f = \frac{3\nabla h}{h}\nabla u - \frac{\lambda N}{h^3}\frac{\partial h}{\partial x}, \tag{2.14}
$$

we can rewrite (2.1)–(2.4) as

$$
-\Delta u \ge f \quad \text{for } (x,y) \in \Omega \tag{2.15}
$$

$$
u \ge 0 \quad \text{for } (x,y) \in \Omega \tag{2.16}
$$

$$
u(-\Delta u - f) = 0 \quad \text{for } (x,y) \in \Omega \tag{2.17}
$$

$$
u = 0 \quad \text{for } (x,y) \in \partial\Omega. \tag{2.18}
$$

LEMMA 2.3. *Under the assumption of Lemma* 2.1, *we have, for any* $2 < p < \infty$,

$$\|u\|_{W^{2,p}(\Omega)} \leq C(\Omega, N, p)\lambda \tag{2.19}$$

$$\|u\|_{W^{2,\infty}(\Omega')} \leq C(\Omega, \Omega', N)\lambda \tag{2.20}$$

$$\|h\|_{W^{1,\infty}(\Omega)} \leq C(\Omega, N) \tag{2.21}$$

$$\|h\|_{W^{2,p}(\Omega')} \leq C(\Omega, \Omega', N) \tag{2.22}$$

where $\Omega' \subset\subset \Omega$, *and the constants in* (2.19)–(2.22) *depend on* k *and* Λ.

Proof. We shall use C to denote various constants depending on Ω and N and use C_p to denote various constants depending on Ω, N, and p.

First let $1 < p < 2$. Then by (2.14)

$$|f|^p \leq C|\nabla h|^p(|\nabla u|^p + \lambda^p). \tag{2.23}$$

Applying Hölder's inequality we get

$$\int_\Omega |f|^p dxdy \leq C\left(\int_\Omega |\nabla h|^{2p/(2-p)} dxdy\right)^{(2-p)/2} \left(\int_\Omega (|\nabla u|^2 + \lambda^2) dxdy\right)^{p/2} \tag{2.24}$$

since $1 < p < 2$, $2 < 2p/(2-p) < \infty$, and by Lemmas 2.1 and 2.2 it then follows that

$$\|f\|_{L^p(\Omega)} \leq C_p\lambda \qquad (1 < p < 2). \tag{2.25}$$

Thus, by L^p estimates for the variational inequality (2.15)–(2.18) give

$$\|u\|_{W^{2,p}(\Omega)} \leq C_p\lambda \qquad (1 < p < 2). \tag{2.26}$$

Using the Sobolev Embedding Theorem we conclude that

$$\|u\|_{W^{1,p}(\Omega)} \leq C_p\lambda \tag{2.27}$$

for any $1 < p < \infty$.

Now from (2.9) and Hölder's inequality it follows that

$$\|\nabla h\|_{L^\infty(\Omega)} \leq C + C\|\nabla u\|_{L^3(\Omega)} \left(\int_{B_K(0)} (s^2 + t^2)^{-3/4} dsdt\right)^{2/3} \tag{2.28}$$

if K is large enough so that $\Omega \subset B_{K/2}(0)$. Using (2.27) it follows that

$$\|\nabla h\|_{L^\infty(\Omega)} \leq C. \tag{2.29}$$

Next, using (2.14), (2.27), and (2.29), we find that

$$\|f\|_{L^p(\Omega)} \leq C_p\lambda \tag{2.30}$$

for any $1 < p < \infty$, and thus by a similar argument as above we obtain the estimate

$$\|u\|_{W^{2,p}(\Omega)} \leq C_p\lambda \tag{2.31}$$

for any $1 < p < \infty$.

To get higher regularity, we differentiate (2.9) using a similar argument as in [5, p. 53–55], and obtain:

$$h_{xx} = 2N^2 + N\int_\Omega \frac{u_{xx}(s,t)dsdt}{\sqrt{(x-s)^2 + (y-t)^2}} - N\int_{\partial\Omega} \frac{u_x(s,t)\cos(\vec{n}, x)}{\sqrt{(x-s)^2 + (y-t)^2}} d\sigma. \tag{2.32}$$

Similar expressions can be derived for h_{xy} and h_{yy}.

By the Sobolev Embedding Theorem and (2.31), it follows that

$$\|u\|_{W^{1,\infty}(\Omega)} \leq C\lambda \tag{2.33}$$

and thus if $(x,y) \in \Omega'$, where $\Omega' \subset\subset \Omega$, then

$$\left| \int_{\partial\Omega} \frac{u_x(s,t)\cos(\vec{n},x)}{\sqrt{(x-s)^2+(y-t)^2}} d\sigma \right| \leq C\lambda. \tag{2.34}$$

Next, applying Young's inequality [5, Lemma 7.12], and using (2.31), (2.34), we conclude that

$$\|h\|_{W^{2,p}(\Omega')} \leq C_p + C_p\|u\|_{W^{2,2}(\Omega)} \leq C_p \quad \text{for any } 2 < p < \infty, \tag{2.35}$$

the constants in (2.35) depend on $dist(\Omega', \partial\Omega)$.

By the Sobolev Embedding Theorem, (2.31), and (2.35), for any $0 < \alpha < 1$ and $\Omega'' \subset\subset \Omega'$,

$$\|h\|_{C^{1,\alpha}(\Omega'')} \leq C' \tag{2.36}$$

and

$$\|u\|_{C^{1,\alpha}(\Omega)} \leq C''\lambda \tag{2.37}$$

where C' depends on α and Ω'' and C'' depends on α. Therefore, using (2.14) it follows that

$$\|f\|_{C^{\beta}(\Omega'')} \leq C'''\lambda \tag{2.38}$$

where C''' depends on β and Ω''. Hence, by elliptic estimates for variational inequalities,

$$\|u\|_{W^{2,\infty}(\Omega''')} \leq C\lambda \tag{2.39}$$

where $\Omega''' \subset\subset \Omega''$, and $0 < \alpha < 1$ (see [4]). □

3. The existence of a solution.

THEOREM 3.1. *Suppose that Ω is a smooth bounded domain, then there exists a solution (u,h) of (1.7)–(1.11) such that*

$$u \in W^{2,p}(\Omega) \cap W^{2,\infty}_{loc}(\Omega), \qquad h \in C^1(\bar{\Omega}) \cap C^{1,\alpha}_{loc}(\Omega) \tag{3.1}$$

for any $2 < p < \infty$ and $0 < \alpha < 1$.

Proof. Take $2 < p < \infty$ (fixed) and let

$$B = W^{1,p}(\Omega) \cap W^{1,2}_0(\Omega). \tag{3.2}$$

For each $u \in B$, define

$$Hu = k + x^2 + y^2 + \int_{\Omega} \frac{u^+(s,t)dsdt}{\sqrt{(x-s)^2+(y-t)^2}}. \tag{3.3}$$

Then, by Hölder's inequality,

$$\|Hu\|_{C^1(\bar{\Omega})} \leq C + C\|u\|_{W^{1,p}(\Omega)} \tag{3.4}$$

$$\|Hu_1 - Hu_2\|_{C^1(\bar{\Omega})} \leq C\|u_1 - u_2\|_{W^{1,p}(\Omega)}. \tag{3.5}$$

Now define Tu to be the solution of (1.7)–(1.10) with $h = Hu$ (it is unique for fixed h).

From (3.4), L^p estimates for the variational inequalities and the compactness of the inclusion $W^{2,p}(\Omega) \to W^{1,p}(\Omega)$, it follows that $T : B \to B$ is compact.

The compactness of T, the uniqueness of the solution of the variational inequality, and (3.5) altogether imply that $T : B \to B$ is continuous.

Next, for any $0 < \sigma < 1$, consider any fixed point u of the operator σT:

$$u = \sigma Tu. \tag{3.6}$$

Notice that since $u/\sigma = Tu$, u is a solution to the following problem:

$$\begin{aligned}
&-\nabla(h^3\nabla u) \geq -\lambda\sigma\frac{\partial h}{\partial x} \quad \text{for } (x,y) \in \Omega \\
&u \geq 0 \quad \text{for } (x,y) \in \Omega \\
&u \cdot \left[-\nabla(h^3\nabla u) + \lambda\sigma\frac{\partial h}{\partial x}\right] = 0 \quad \text{for } (x,y) \in \Omega \\
&u = 0 \quad \text{for } (x,y) \in \partial\Omega
\end{aligned}$$

and

$$h(x,y) = k + x^2 + y^2 + \int_\Omega \frac{u(s,t)dsdt}{\sqrt{(x-s)^2 + (y-t)^2}}. \tag{3.7}$$

Hence by Lemma 2.3 (with $N = 1$),

$$\|u\|_{W^{1,p}(\Omega)} \leq C \tag{3.8}$$

where C is a constant independent of σ.

From this fact and the previous properties of T it follows that the Laray–Schauder fixed point theorem [5, Thm. 10.3] can be applied. Thus, there is a fixed point u for T, that is, there exists a solution (u,h) to the problem (1.7)–(1.11).

Finally by Lemma 2.3,

$$u \in W^{2,\infty}_{loc}(\Omega), \qquad h \in C^{1,\alpha}_{loc}(\Omega), \tag{3.9}$$

and the theorem is proved. □

Remark. If the domain Ω is symmetric with respect to the x-axis, then we may take in (3.2)

$$B = W^{1,p}(\Omega) \cap W^{1,2}_0(\Omega) \cap \{u | u(x,y) = u(x,-y)\} \tag{3.10}$$

and the preceding argument shows that there exists a solution symmetric with respect to y.

4. Estimate on the support. As a simple result of (2.33), we have the following theorem for fixed Ω.

THEOREM 4.1. *For any $\epsilon > 0$, there exists $\lambda_* > 0$ such that if $0 < \lambda < \lambda_*$, then*

$$u(x,y) > 0 \quad \text{for } x < -\epsilon, \ (x,y) \in \Omega. \tag{4.1}$$

Proof. From (2.9) (with N=1) it follows that

$$\begin{aligned} h_x &= 2x + \int_\Omega \frac{u_x(s,t)dsdt}{\sqrt{(x-s)^2+(y-t)^2}} \\ &\equiv 2x + I. \end{aligned} \tag{4.2}$$

By (2.33)

$$\begin{aligned} I &\le C\lambda \int_\Omega \frac{dsdt}{\sqrt{(x-s)^2+(y-t)^2}} \\ &\le C\lambda; \end{aligned} \tag{4.3}$$

hence, for $x < -\epsilon$,

$$h_x < -2\epsilon + C\lambda < 0 \quad \text{for } x < -\epsilon, \ \lambda < \lambda_*, \tag{4.4}$$

provided λ_* is small enough, and hence $u(x,y) > 0$ for $x < -\epsilon$. □

If λ is not small, a similar result still holds if the domain is "large" enough, i.e., $\Omega = B_M(0)$ in (1.7)–(1.11), where M is large. We shall prove the following theorem.

THEOREM 4.2. *For any $\epsilon > 0$, there exists a $K > 0$ such that if $M > K$, then*

$$u(x,y) > 0 \quad \text{for } -M < x < -\epsilon M, \ (x,y) \in B_M(0). \tag{4.5}$$

To prove this theorem, we start with a scaling:

$$u_M(x,y) = u(Mx, My) \quad \text{for } x^2 + y^2 \le 1 \tag{4.6}$$

$$h_M(x,y) = h(Mx, My) \quad \text{for } x^2 + y^2 \le 1. \tag{4.7}$$

A simple calculation shows

$$-\nabla(h_M^3 \nabla u_M) \ge -\lambda M \frac{\partial h_M}{\partial x} \quad \text{for } (x,y) \in B_1 \tag{4.8}$$

and

$$h_M(x,y) = k + M^2(x^2+y^2) + M \int_{B_1} \frac{u(s,t)dsdt}{\sqrt{(x-s)^2+(y-t)^2}}. \tag{4.9}$$

This shows that (u_M, h_M) satisfies (2.1)–(2.5) with $\Omega = B_1(0)$.

It clearly suffices to show that

$$u_M(x,y) > 0 \quad \text{for } -1 < x < -\epsilon \tag{4.10}$$

for M large enough; for simplicity, we drop the subscripts M from u_M and h_M.

Since $h \geq \epsilon^2 M^2$ if $x^2 + y^2 \geq \epsilon^2$, we get from (2.6),

$$\int_{B_1 \setminus B_\epsilon} |\nabla u|^2 dxdy \leq \frac{\lambda^2 \pi}{k\epsilon^6} M^{-4}. \tag{4.11}$$

We shall use C to denote various constants independent of M (although they may depend on ϵ, k, and Λ).

We are going to carry out a proof similar to that in Lemma 2.3, but this time we shall use the fact that $h \geq \epsilon^2 M^2$ if $x^2 + y^2 \geq \epsilon^2$ to find a better estimate on h when $x^2 + y^2 \geq (4\epsilon)^2$.

By Hölder's inequality, for $1 < p < 2$

$$\left\| \frac{\nabla h}{h} \nabla u \right\|_{L^p(B_1 \setminus B_\epsilon)} \leq \frac{1}{\epsilon^2 M^2} \|\nabla h\|_{L^{2p/(2-p)}(B_1 \setminus B_\epsilon)} \|\nabla u\|_{L^2(B_1 \setminus B_\epsilon)}. \tag{4.12}$$

By using (4.11) to estimate ∇u and Lemma 2.2 to estimate ∇h, we get

$$\left\| \frac{\nabla h}{h} \nabla u \right\|_{L^p(B_1 \setminus B_\epsilon)} \leq \frac{1}{\epsilon^2 M^2} (C_p M^2)(CM^{-2}) \leq \frac{C_p}{M^2}. \tag{4.13}$$

By Lemma 2.2, also

$$\begin{aligned} \left\| \frac{\lambda M}{h^3} \frac{\partial h}{\partial x} \right\|_{L^p(B_1 \setminus B_\epsilon)} &\leq \frac{\lambda M}{(\epsilon^2 M^2)^3} \|\nabla h\|_{L^p(B_1)} \\ &\leq (CM^{-5})(C_p M^2) \leq C_p M^{-3}. \end{aligned} \tag{4.14}$$

From (2.14), (4.13), and (4.14), it now follows that

$$\|f\|_{L^p(B_1 \setminus B_\epsilon)} \leq C_p M^{-2} \qquad (1 < p < 2). \tag{4.15}$$

Thus we prove the following lemma.

LEMMA 4.3. *For f defined in* (2.14) (*with $N = M$*),

$$\|f\|_{L^p(B_1 \setminus B_\epsilon)} \leq C_p M^{-2} \quad (1 < p < 2). \tag{4.16}$$

□

Notice that u satisfies the variational inequality

$$-\Delta u \geq f, \quad u \geq 0, \quad u(-\Delta u - f) = 0 \quad \text{in } B_1 \tag{4.17}$$

$$u = 0 \quad \text{on } \partial B_1 \tag{4.18}$$

where f is given by (2.14) (with $N = M$); by Lemmas 4.3, (4.11), and the Poincaré inequality,

$$\|v\|_{L^2(B_1 \setminus B_\epsilon)} \leq \|\nabla v\|_{L^2(B_1 \setminus B_\epsilon)} \quad \text{for } v \in W_0^{1,2}(B_1), \tag{4.19}$$

we get:

$$\|u\|_{L^2(B_1 \setminus B_\epsilon)} \leq CM^{-2} \qquad (1 < p < 2) \tag{4.20}$$

$$\|\nabla u\|_{L^2(B_1 \setminus B_\epsilon)} \leq CM^{-2} \qquad (1 < p < 2) \tag{4.21}$$

$$\|f\|_{L^p(B_1 \setminus B_\epsilon)} \leq C_p M^{-2} \qquad (1 < p < 2). \tag{4.22}$$

LEMMA 4.4. *For* $1 < p < 2$

$$\|u\|_{W^{2,p}(B_1 \setminus B_{2\epsilon})} \le C_p M^{-2} \qquad (1 < p < 2), \tag{4.23}$$

and hence, by the Sobolev Embedding Theorem,

$$\|u\|_{W^{1,p}(B_1 \setminus B_{3\epsilon})} \le C_p M^{-2} \qquad (1 < p < \infty). \tag{4.24}$$

Proof. Take a cutoff function $\zeta \in C^\infty$ so that

$$\begin{aligned} \zeta &= 1 \quad \text{for } 2\epsilon \le \sqrt{x^2+y^2} \le 1 \\ &= 0 \quad \text{for } \sqrt{x^2+y^2} \le \epsilon \end{aligned}$$

and $0 \le \zeta \le 1$. Let $w = \zeta u$. Then

$$\Delta w = u\Delta\zeta + 2\nabla\zeta\nabla u + \zeta\Delta u, \tag{4.25}$$

and w satisfies the variational inequality

$$\begin{aligned} &-\Delta w \ge F \quad \text{for } \epsilon < \sqrt{x^2+y^2} < 1 \\ &w \ge 0 \quad \text{for } \epsilon < \sqrt{x^2+y^2} < 1 \\ &w(-\Delta w - F) = 0 \quad \text{for } \epsilon < \sqrt{x^2+y^2} < 1 \\ &w = 0 \quad \text{for } \sqrt{x^2+y^2} = \epsilon,\ \sqrt{x^2+y^2} = 1 \end{aligned}$$

where

$$F = -u\Delta\zeta - 2\nabla\zeta\nabla u - \zeta f. \tag{4.26}$$

By (4.20)–(4.22), we get

$$\|F\|_{L^p(B_1 \setminus B_\epsilon)} \le C_p M^{-2} \qquad (1 < p < 2). \tag{4.27}$$

Thus by L^p estimates for the variational inequality,

$$\|w\|_{W^{2,p}(B_1 \setminus B_\epsilon)} \le C_p M^{-2} \qquad (1 < p < 2) \tag{4.28}$$

and (4.23) follows. □

Next, we prove the following lemma.

LEMMA 4.5. *There exists a constant* C *such that*

$$\int_{B_1} u(x,y)dxdy \le CM^{-1/5} \tag{4.29}$$

uniformly for large M.

Proof. If (4.29) is not true, then there exists a sequence $M_n \to \infty$ such that

$$\int_{B_1} u_{M_n}(x,y)dxdy > nM_n^{-1/5}. \tag{4.30}$$

Thus

$$\int_{B_1} \frac{u_{M_n}(s,t)}{\sqrt{(x-s)^2+(y-t)^2}} dsdt \ge \frac{1}{2}\int_{B_1} u_{M_n}(x,y)dxdy > \frac{1}{2}nM_n^{-1/5}; \tag{4.31}$$

hence

$$h_{M_n} \geq \frac{1}{2} n M_n^{1-(1/5)} > \frac{n}{2} M_n^{4/5}. \tag{4.32}$$

By Lemma 2.1,

$$\int_{B_1} \left(\frac{n}{2} M_n^{4/5}\right)^3 |\nabla u_{M_n}|^2 dxdy \leq C M_n^2, \tag{4.33}$$

and thus

$$\int_{B_1} |\nabla u_{M_n}|^2 dxdy \leq C n^{-3} M_n^{2-(12/5)} = C n^{-3} M_n^{-2/5}. \tag{4.34}$$

Using Hölder's inequality and (4.19) with $\epsilon \to 0$, we get

$$\begin{aligned} \int_{B_1} u_{M_n}(x,y) dxdy &\leq \pi^{1/2} \|u_{M_n}\|_{L^2(B_1)} \\ &\leq \pi^{1/2} \|\nabla u_{M_n}\|_{L^2(B_1)} \\ &\leq C n^{-3/2} M_n^{-1/5}. \end{aligned} \tag{4.35}$$

From (4.30) and (4.35), it follows that

$$n M_n^{-1/5} < C n^{-3/2} M_n^{-1/5} \tag{4.36}$$

or

$$n < C n^{-3/2} \tag{4.37}$$

which is a contradiction. □

Proof of Theorem 4.2. From (2.9)

$$h_x = 2xM^2 + MI \tag{4.38}$$

where

$$I = \int_{B_1} \frac{u_x(s,t)}{\sqrt{(x-s)^2 + (y-t)^2}} dsdt. \tag{4.39}$$

Take a cutoff function $\zeta \in C^\infty$ such that

$$\begin{aligned} \zeta &= 1 \quad \text{for } \sqrt{x^2+y^2} \geq 3\epsilon \\ &= 0 \quad \text{for } \sqrt{x^2+y^2} \leq 2\epsilon \end{aligned} \tag{4.40}$$

and $0 \leq \zeta \leq 1$. Then

$$\begin{aligned} I &= \int_{B_1} \frac{(u\zeta + u(1-\zeta))_x(s,t)}{\sqrt{(x-s)^2 + (y-t)^2}} dsdt \\ &= \int_{B_1} \frac{(u\zeta)_x(s,t)}{\sqrt{(x-s)^2 + (y-t)^2}} dsdt + \int_{B_1} \frac{(u(1-\zeta))_x(s,t)}{\sqrt{(x-s)^2 + (y-t)^2}} dsdt \\ &\equiv J_1 + J_2. \end{aligned} \tag{4.41}$$

Assume that $(x,y) \in B_1 \setminus B_{4\epsilon}$. Then

$$\sqrt{(x-s)^2+(y-t)^2} \geq \epsilon \quad \text{for } (s,t) \in B_{3\epsilon}. \tag{4.42}$$

Since $u(1-\zeta)=0$ on $\partial B_{3\epsilon}$, no boundary term will appear when we use integration by parts for J_2. Hence

$$\begin{aligned} J_2 &= \int_{B_{3\epsilon}} \frac{(u(1-\zeta))_x(s,t)}{\sqrt{(x-s)^2+(y-t)^2}} dsdt \\ &= -\int_{B_{3\epsilon}} u(1-\zeta)\left(\frac{\partial}{\partial s}\frac{1}{\sqrt{(x-s)^2+(y-t)^2}}\right) dsdt \end{aligned} \tag{4.43}$$

and thus

$$\begin{aligned} |J_2| &\leq C\int_{B_{3\epsilon}} |u| dsdt \\ &\leq CM^{-1/5}, \end{aligned} \tag{4.44}$$

the last inequality is obtained by Lemma 4.5.

By Hölder's inequality

$$\begin{aligned} \|J_1\|_{L^\infty(B_1)} &\leq \|(u\zeta)_x\|_{L^3(B_1\setminus B_{2\epsilon})} \left[\sup_{(x,y)\in B_1} \left\| \frac{1}{\sqrt{(x-\cdot)^2+(y-\cdot)^2}} \right\|_{L^{3/2}(B_1)}\right] \\ &\leq C\|\nabla(u\zeta)\|_{L^3(B_1\setminus B_{2\epsilon})} \end{aligned} \tag{4.45}$$

where the last inequality is obtained by using (4.24).

Thus, by (4.44), (4.45), and (4.21), for $(x,y) \in B_1 \setminus B_{4\epsilon}$

$$|I| \leq |J_1| + |J_2| \leq CM^{-1/5}, \tag{4.46}$$

and thus, for $x \leq -4\epsilon$, we have

$$\begin{aligned} h_x &\leq 2xM^2 + CM^{1-(1/5)} \\ &\leq -8\epsilon M^2 + CM^{2/5} < 0 \end{aligned} \tag{4.47}$$

if $M > K(\epsilon)$, and hence $u(x,y) > 0$ for $x < -4\epsilon$. □

5. Uniqueness. In this section we prove uniqueness provided λ is small. The estimates that we obtained in §§2 and 3 are uniform for λ, that is, the constants C depend only on Λ. Ω will be a fixed smooth domain.

THEOREM 5.1. *There exists $\lambda_1 > 0$ such that the solution of* (1.7)-(1.11) *is unique if $0 \leq \lambda \leq \lambda_1$, where λ_1 depends on Ω and k.*

Proof. We shall use C to denote constants that do not depend on λ.

If (u,h),$(\tilde{u},\tilde{h})$ are two solutions, then we have

$$[-\nabla(h^3\nabla u) + \lambda h_x](\tilde{u}-u) \geq 0 \tag{5.1}$$

$$[-\nabla(\tilde{h}^3\nabla \tilde{u}) + \lambda \tilde{h}_x](u-\tilde{u}) \geq 0. \tag{5.2}$$

Thus

$$\int_\Omega h^3\nabla u\nabla(\tilde{u}-u) \geq \lambda\int_\Omega h(\tilde{u}-u)_x \tag{5.3}$$

$$\int_\Omega \tilde{h}^3\nabla\tilde{u}\nabla(u-\tilde{u}) \geq \lambda\int_\Omega \tilde{h}(u-\tilde{u})_x; \tag{5.4}$$

hence

$$\int_\Omega \{h^3|\nabla(u-\tilde{u})|^2 + (\tilde{h}^3 - h^3)\nabla\tilde{u}\nabla(\tilde{u}-u)\} \leq \lambda \int_\Omega |\nabla(u-\tilde{u})||h-\tilde{h}|. \tag{5.5}$$

Using (2.19) and (2.21), we get

$$\begin{aligned} k^3 \int_\Omega |\nabla(u-\tilde{u})|^2 &\leq C\lambda \int_\Omega |\nabla(u-\tilde{u})||h-\tilde{h}| \\ &\leq C\lambda \|\nabla(u-\tilde{u})\|_{L^2(\Omega)} \|h-\tilde{h}\|_{L^2(\Omega)}. \end{aligned} \tag{5.6}$$

By Young's inequality [5, Lemma 7.12] with $p = q = 2$, we get

$$\|h-\tilde{h}\|_{L^2(\Omega)} \leq C\|u-\tilde{u}\|_{L^2(\Omega)}. \tag{5.7}$$

Now (5.6) and (5.7), together with Poincaré's inequality give:

$$\int_\Omega |\nabla(u-\tilde{u})|^2 \leq C\lambda \|\nabla(u-\tilde{u})\|_{L^2(\Omega)} \|h-\tilde{h}\|_{L^2(\Omega)} \leq C\lambda \int_\Omega |\nabla(u-\tilde{u})|^2. \tag{5.8}$$

Thus, if λ is small, then

$$\int_\Omega |\nabla(u-\tilde{u})|^2 = 0 \tag{5.9}$$

which implies that $u = \tilde{u}$. □

6. The case $\mu = \mu_0 e^{\alpha u}$. In the previous sections we studied the case when the viscosity is constant. We now study the case when the viscosity is given by (1.13). We shall extend all the previous results to this case provided α is small.

Using the transformation

$$w = 1 - e^{-\alpha u}, \tag{6.1}$$

we obtain from (1.1)–(1.5) the variational inequality:

$$-\nabla(h^3\nabla w) \geq -\lambda\alpha h_x \quad \text{for } (x,y) \in \Omega \tag{6.2}$$

$$w \geq 0 \quad \text{for } (x,y) \in \Omega \tag{6.3}$$

$$w[-\nabla(h^3\nabla w) + \lambda\alpha h_x] = 0 \quad \text{for } (x,y) \in \Omega \tag{6.4}$$

$$w = 0 \quad \text{for } (x,y) \in \partial\Omega \tag{6.5}$$

and

$$h(x,y) = k + x^2 + y^2 + \int_\Omega \frac{1}{\alpha}\left(\log \frac{1}{1-w(s,t)}\right) \frac{dsdt}{\sqrt{(x-s)^2 + (y-t)^2}}. \tag{6.6}$$

The main difficulty is to show that $1 - w$ stays uniformly positive. To overcome this difficulty, let us fix $\epsilon \in (0,1)$ and take a cutoff function ζ such that $\zeta \in C^\infty$, $0 \leq \zeta \leq 1$ and

$$\begin{aligned} \zeta(w) &= 1 \quad \text{for } w \leq 1-\epsilon \\ &= 0 \quad \text{for } w \geq 1-\epsilon/2. \end{aligned}$$

Define

$$G(w) = \zeta(w) \log \frac{1}{1-w} \quad \text{for } 0 \le w < \infty; \tag{6.7}$$

then $G \in C^\infty$. Let us take, for instance, $\epsilon = \frac{1}{2}$. Next, instead of (6.6), take

$$h(x,y) = k + x^2 + y^2 + \int_\Omega \frac{1}{\alpha} G(w(s,t)) \frac{dsdt}{\sqrt{(x-s)^2 + (y-t)^2}}. \tag{6.8}$$

Let us consider the system (6.2)–(6.5) and (6.8). From (6.2)–(6.5), it follows that

$$\int_\Omega |\nabla w|^2 dxdy \le \frac{\lambda^2 \alpha^2 |\Omega|}{k^4}. \tag{6.9}$$

Differentiating (6.8), we get

$$\nabla h = 2(x,y) + \int_\Omega G'(w(s,t)) \frac{\nabla w(s,t)}{\alpha} \frac{dsdt}{\sqrt{(x-s)^2 + (y-t)^2}}. \tag{6.10}$$

Note that $|G'(w(s,t))|$ is bounded uniformly, and thus, by Young's inequality [5, Lemma 7.12], we get, for $2 < p < \infty$,

$$\|\nabla h\|_{L^p(\Omega)} \le C + C \|\frac{1}{\alpha} \nabla w\|_{L^2(\Omega)}; \tag{6.11}$$

by (6.9)

$$\|\nabla h\|_{L^p(\Omega)} \le C \tag{6.12}$$

where the constant C is independent of α. Thus, if

$$f = \frac{3\nabla h}{h} \nabla w - \frac{\lambda \alpha}{h^3} \frac{\partial h}{\partial x} \tag{6.13}$$

then, as in the proof of Lemma 2.3 (using (6.9)),

$$\|f\|_{L^p(\Omega)} \le C_p \alpha \lambda \qquad (1 < p < 2). \tag{6.14}$$

From (6.9) and (6.14), it follows by using the L^p estimates for the variational inequality that

$$\|w\|_{W^{2,p}(\Omega)} \le C_p \alpha \lambda \qquad (1 < p < 2) \tag{6.15}$$

Applying the Sobolev Embedding Theorem to (6.15), we obtain

$$w \le C \alpha \lambda \le 1 - \epsilon = \frac{1}{2} \tag{6.16}$$

provided $\alpha\lambda$ is small, and then the expressions (6.8) and (6.6) coincide. The existence, regularity, and uniqueness of the solution now follows as before. Let us summarize the results as follows.

THEOREM 6.1. *There exists a constant $c = c(k, \Omega) > 0$ such that if $0 < \alpha\lambda < c$, then there exists a solution (u, h) such that*

$$u \in W^{2,p}(\Omega) \cap W^{2,\infty}_{loc}(\Omega), \quad h \in C^1(\bar{\Omega}) \cap C^{1,\alpha}_{loc}(\Omega). \tag{6.17}$$

Moreover, there exists a constant $\lambda_1 = \lambda_1(k, \Omega) > 0$ such that if $0 < \alpha\lambda < c$, $0 < \lambda < \lambda_1$, then the solution (u, h) is unique.

7. The one-dimensional problem. Consider the one-dimensional variational inequality:

$$-\left(h^3\frac{u'}{e^{\alpha u}}\right)' \geq -\lambda h' \quad \text{for } x \in [-M, M] \tag{7.1}$$

$$u \geq 0 \quad \text{for } x \in [-M, M] \tag{7.2}$$

$$u \cdot \left[-\left(h^3\frac{u'}{e^{\alpha u}}\right)' + \lambda h'\right] = 0 \quad \text{for } x \in [-M, M] \tag{7.3}$$

$$u(\pm M) = 0 \tag{7.4}$$

and

$$h = k + x^2 + \int_{-M}^{M} u(s) \log \frac{2M}{|x-s|} ds \tag{7.5}$$

where M, λ, and k are positive constants; and α is a positive constant which shall be assumed to be sufficiently small in proving the existence of the solution; both α and λ will be assumed to be sufficiently small in proving the uniqueness of the solution.

Introducing the transformation

$$w = 1 - e^{-\alpha u}, \tag{7.6}$$

as before, we obtain

$$u = \frac{1}{\alpha} \log \frac{1}{1-w}, \tag{7.7}$$

and (7.1)–(7.5) is transformed into the following problem:

$$-(h^3 w')' \geq -\lambda\alpha h' \quad \text{for } x \in [-M, M] \tag{7.8}$$

$$w \geq 0 \quad \text{for } x \in [-M, M] \tag{7.9}$$

$$w[-(h^3 w')' + \lambda\alpha h'] = 0 \quad \text{for } x \in [-M, M] \tag{7.10}$$

$$w(\pm M) = 0 \tag{7.11}$$

and

$$h = k + x^2 + \frac{1}{\alpha} \int_{-M}^{M} \left(\log \frac{1}{1-w(s)}\right) \log \frac{2M}{|x-s|} ds. \tag{7.12}$$

As we did in §6, instead of (7.12), we consider

$$h = k + x^2 + \frac{1}{\alpha} \int_{-M}^{M} (G(w(s))) \log \frac{2M}{|x-s|} ds \tag{7.13}$$

where $G(w)$ is defined in (6.7). The system (7.1)–(7.5) is equivalent to (7.8)–(7.11) and (7.13) if we can establish the bound:

$$w \leq 1 - \epsilon \tag{7.14}$$

but this follows, for small α, from the following lemma.

LEMMA 7.1. *Suppose that* $\lambda \leq \Lambda$. *If* w *satisfies* (7.8)–(7.11), *then*

$$|w(y)| \leq 2M\frac{\lambda\alpha}{k^2} \quad \text{for } |y| \leq M. \tag{7.15}$$

Proof. Integrating (7.10) over $[-M, M]$, we get

$$\begin{aligned} \int_{-M}^{M} h^3|w'|^2 &= -\lambda\alpha\int_{-M}^{M} h'w = \lambda\alpha\int_{-M}^{M} hw' \\ &\leq \lambda\alpha\left(\int_{-M}^{M} h^2|w'|^2\right)^{1/2}(2M)^{1/2} \\ &\leq \frac{\lambda\alpha}{k^{1/2}}(2M)^{1/2}\left(\int_{-M}^{M} h^3|w'|^2\right)^{1/2}, \end{aligned} \tag{7.16}$$

and hence

$$\int_{-M}^{M} h^3|w'|^2 \leq \frac{\lambda^2\alpha^2}{k}(2M). \tag{7.17}$$

It follows that

$$\|w'\|_{L^2[-M,M]} \leq \frac{\lambda\alpha}{k^2}(2M)^{1/2}, \tag{7.18}$$

so that

$$|w(y)| \leq \int_{-M}^{M} |w'(s)|ds \leq (2M)^{1/2}\|w'\|_{L^2[-M,M]}. \tag{7.19}$$

Using (7.18), (7.15) follows. □

Lemma 7.1 tells us that if we choose, for instance, $\epsilon = \frac{1}{2}$, then there exists an $A > 0$ (we may take $A = (k^2/2M)(1-\epsilon) = k^2/4M$) such that

$$w(y) \leq 1 - \epsilon = \frac{1}{2} \quad \text{for } 0 < \alpha\lambda < A, \tag{7.20}$$

and hence (7.12) and (7.13) coincide. Next we shall use constants C to denote various constants depending only on M, k, A, and Λ. Since we are going to use the same method as in §3 to prove existence, care has to be taken so that the constants will not change when we replace λ by $\sigma\lambda$ $(0 \leq \sigma \leq 1)$.

Note that,

$$\lim_{\delta\to 0}\frac{1}{\delta}\log\frac{1}{1-2M\delta/k^2} = \frac{2M}{k^2}. \tag{7.21}$$

and, consequently, by (7.15),

$$\sup_{|s|\leq M}\left|\frac{1}{\alpha}\log\frac{1}{1-w(s)}\right| \leq \lambda\left(\frac{1}{\alpha\lambda}\log\frac{1}{1-2M\alpha\lambda/k^2}\right) \leq C\lambda. \tag{7.22}$$

By (7.12),

$$h \leq C + C\int_{-M}^{M}\log\frac{2M}{|x-s|}ds \leq C. \tag{7.23}$$

To derive an estimate for $w'(x)$ for all x, it suffices to derive an estimate for $w'(x)$ for every interval $[a,b]$ such that $w(a) = w(b) = 0$ and $w(x) > 0$ for $x \in (a,b)$.

By Rolle's theorem, there is a number $c \in (a,b)$ such that $w'(c) = 0$. From (7.10), it follows that

$$h^3(x)w'(x) = \lambda\alpha(h(x) - h(c)). \tag{7.24}$$

And hence, by (7.23),

$$\frac{1}{\alpha}|w'(x)| \le \lambda\left|\frac{h(x) - h(c)}{h^3(x)}\right| \le C\lambda. \tag{7.25}$$

Differentiating (7.13) (which is now the same as (7.12)), we get

$$h'(x) = 2x + \int_{-M}^{M} \frac{1}{\alpha}\frac{w'(s)}{1 - w(s)} \log\frac{2M}{|x - s|}ds, \tag{7.26}$$

and, by (7.20), (7.25),

$$|h'(x)| \le 2M + C\int_{-M}^{M} \log\frac{2M}{|x - s|}ds \le C. \tag{7.27}$$

By (7.10),

$$h^3w'' + 3h^2h'w' = \lambda\alpha h' \quad \text{if } w(x) > 0, \tag{7.28}$$

and thus, by (7.25), (7.27),

$$\sup_{|x|\le M} \frac{1}{\alpha}|w''(x)| \le C\lambda. \tag{7.29}$$

Using these estimates, we can prove the existence of the solution by using the same method as in §3. In the case $\alpha = 0$, we may work with u instead of w. We have proved the following theorem.

THEOREM 7.2. *If α is sufficiently small such that*

$$\frac{2M\lambda\alpha}{k^2} < 1, \tag{7.30}$$

then there exists a solution to (7.1)–(7.5). □

Remark. This condition on α is much more precise than that derived in Theorem 6.1.

Next, we prove that the free boundary consists of at most one free boundary point.

THEOREM 7.3. *There exists $\lambda_0 = \lambda_0(k, M)$ such that if $0 \le \alpha\lambda \le A$, $0 \le \lambda \le \lambda_0$, then there is at most one free boundary point.*

Proof. We can rewrite h' as:

$$h'(x) = 2x + \int_{-M-x}^{M-x} \frac{1}{\alpha}\frac{w'(t + s)}{1 - w(t + s)} \log\frac{2M}{|t|}dt, \tag{7.31}$$

then

$$\begin{aligned} h''(x) &= 2 + \int_{-M}^{M} \frac{1}{\alpha}\left[\frac{w''(s)}{1-w(s)} + \frac{(w'(s))^2}{(1-w(s))^2}\right] \log \frac{2M}{|x-s|} ds \\ &\quad - \frac{1}{\alpha} w'(M) \log \frac{2M}{M-x} + \frac{1}{\alpha} w'(-M) \log \frac{2M}{M+x} \\ &= 2 - I. \end{aligned}$$

By (7.25), (7.29), we get

$$h''(x) \geq 2 - \lambda C_1 \quad \text{for } |x| \leq \frac{M}{2} \tag{7.32}$$

so that if λ is small

$$h''(x) > 0 \quad \text{for } |x| \leq \frac{M}{2}. \tag{7.33}$$

By (7.25), (7.26), if λ is small enough, then

$$h'(x) \geq M - \lambda C_2 > 0 \quad \text{for } x \geq \frac{M}{2} \tag{7.34}$$

$$h'(x) \leq -M + \lambda C_2 < 0 \quad \text{for } x \leq -\frac{M}{2}. \tag{7.35}$$

Take λ_0 so that (7.33)–(7.35) hold for $\lambda \leq \lambda_0$. Then $h'(x)$ has only one zero in $[-M, M]$, say, at $x = d$. Then

$$h'(x) < 0 \quad \text{for } -M \leq x < d \tag{7.36}$$

$$h'(x) > 0 \quad \text{for } d < x \leq M. \tag{7.37}$$

Suppose now that b is the first free boundary point, i.e.,

$$w(x) > 0 \quad \text{for } -M < x < b \tag{7.38}$$

$$w(b) = 0 \quad \text{for } b < M. \tag{7.39}$$

By regularity, $w'(b) = 0$. By (7.36), $w > 0$ on $(-M, d)$. Hence $b \geq d$.

Let us define

$$\begin{aligned} \tilde{w}(x) &= w(x) \quad \text{for } -M \leq x < b \\ &= 0 \quad \text{for } b \leq x \leq M. \end{aligned} \tag{7.40}$$

Since $h'(x) > 0$ for $x > b$, $\tilde{w}$ is also a solution of (7.8)–(7.11). Since the solution is unique (for fixed h), we have $w = \tilde{w}$, and hence b is the only free boundary point. □

We also have uniqueness if λ is small.

THEOREM 7.4. *There exists a number $\lambda_1 > 0$ so that the solution (w, h) to the problem (7.8)–(7.12) is unique provided that $0 \leq \lambda \leq \lambda_1$.*

Proof. The proof is essentially the same as that in Theorem 5.1. It is necessary only to check that

$$\|h - \tilde{h}\|_{L^2[-M,M]} \leq C\lambda \|u - \tilde{u}\|_{L^2[-M,M]} \tag{7.41}$$

which is obviously true. □

Remark. Several problems remain open: the uniqueness of the solution without assuming that λ is small, the shape of the free boundary in the two-dimensional case, and the existence of the solution when $\mu = \mu_0 e^{\alpha u}$ without assuming α small.

Acknowledgments. The problem considered in this paper was suggested by Edward Bissett from General Motor Research Laboratories. The author thanks Professor Avner Friedman for his direction and help while working on this paper.

REFERENCES

[1] E. J. BISSETT AND D. W. GLANDER, *A highly accurate approach that resolves the pressure spike of elastohydrodynamic lubrication,* Trans. ASME, J. Tribiology, 110 (1988), pp. 241–246.

[2] G. CAPRIZ AND G. CIMMATI, *Free boundary problems in the theory of hydrodynamic lubrication: a survey,* Free Boundary Problem: Theory and Application. Vol. II, Pitman, London, 1983, pp. 613–635.

[3] D. DOWSON AND G. R. HIGGINSON, *Elastohydrodynamic Lubrication,* Pergamon Press, Oxford, 1966.

[4] A. FRIEDMAN, *Variational Principles and Free Boundary Problems,* John Wiley and Sons, New York, 1982.

[5] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equation of Second Order,* Springer-Verlag, Berlin, 1977.

[6] S. M. ROHDE AND K. P. OH, *A unified treatment of thick and thin film elastohydrodynamic problems by using higher-order element method,* Proc. Roy. Soc. London, Ser. A, 343 (1975), pp. 315–331.

[7] S. R. WU, *A penalty formulation and numerical approximation of the Reynolds–Hertz problem of elastohydrodynamic lubrication,* Internat. J. Engrg. Sci., 24 (1986), pp. 1001–1013.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 37-52, January 1990
© 1990 Society for Industrial and Applied Mathematics
003

SUR UNE CLASSE DE FONCTIONNELLES NON CONVEXES ET APPLICATIONS*

RABAH TAHRAOUI†

Abstract. Various questions of physical or mechanical nature are frequently solved by a variational approach. In many situations, minimizers for the total energy associated with the system are being sought. This energy is represented by an integral functional

$$J(v)=\int_\Omega g(x, v, A\,v)\,dx$$

where v is a vectorial function mapping a bounded open subset Ω of $\mathbb{R}^n$ into $\mathbb{R}^m$, and A is a differential operator.

The goal of this paper is closely related to the study of the elastostatic equilibrium for materials whose constitutive laws are nonlinear. In some realistic situations, the shape of the body, the nature of the deformation, or some symmetric arguments require Ω to be an annulus or a disc.

For example, let B be an isotropic and homogeneous body occupying a reference configuration $\mathcal{R}$ before deformation and $\mathcal{R}'$ after deformation

$$x=(x_1, x_2, x_3)\in\mathcal{R}\to x'=x+u(x)\in\mathcal{R}'$$

where $u(x)$ is the displacement function. It is assumed that the deformation is of the form

$$x=(r, \theta, z)\to x'=(u(r), \theta+v(r), z+w(r))$$

where $r=((x_1^2+x_2^2)^{1/2}z=x_3$ and $\theta=\operatorname{arctg}(x_2/x_1)$. The associated energy can be represented by

$$J_1(u)=\int_\Omega g_1(U(|y|, u, \nabla u))\,dy+\lambda\int_\Omega h_1(|y|, u)\,dy$$

where U is a certain function that will be specified later. The function J_1 is, in general, nonconvex; then the direct method of calculus of variations is not applicable. The lack of weak lower sequential semicontinuity does not make it possible to tend toward the limit in minimizing sequences. Despite this, some existence and regularity results are proved by relaxation techniques.

Key words. elasticité non linéaire, fonctionnelle énergie, non convexe, calcul des variations, relaxation, minimisation

AMS(MOS) subject classifications. 73G50, 73B05, 58E30, 49A50, 58G20

1. Introduction.

L'étude de nombreux phénomènes physiques conduit à la recherche de fonctions minimisant l'énergie interne du système. Cette énergie est en général donnée par des fonctionnelles du type

$$J(v)=\int_\Omega g(x, v, A\,v)\,dx,$$

où $v=(v_1, v_2, \cdots, v_m)$ désigne une fonction définie sur l'ouvert borné Ω de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}^m$, et A est un opérateur différentiel. Une motivation de ce travail est l'étude de l'équilibre élastostatique de certains matériaux dont la loi de comportement est non linéaire (cf. [6], [19], [15], [7]). La géométrie de certains corps et la nature de leur déformation nous imposent de travailler avec Ω à symétrie radiale, i.e., Ω est soit une couronne soit un disque de $\mathbb{R}^2$, de centre zéro: il s'agit, par exemple, de l'étude

* Received by the editors March 1, 1988; accepted for publication (in revised form) February 2, 1989.
† Université de Paris-Sud et Centre, Bâtiment 425, Nationale de Recherche Scientifique, Laboratoire d'Analyse Numérique d'Orsay, Orsay 91405, France and Université de Picardie, U.F.R. Cedex Mathematique et Informatique, Amiens, France.

du déplacement $u=(u_1, u_2, u_3)$ d'un corps B élastique, isotrope, et homogène qui occupe les positions $\mathcal{R}=\Omega\times]-L, +L[$ et $\mathcal{R}'$, respectivement, avant et après déformation:

$$x=(x_1, x_2, x_3)\in\mathcal{R}\to x'=x+u(x)\in\mathcal{R}',$$

on suppose que cette déformation est de la forme

$$x=(r, \theta, z)\to x'=(u(r), \theta+v(r), z+w(r))$$

avec $r=(x_1^2+x_2^2)^{1/2}$ et $z=x_3$; θ désigne l'angle polaire du vecteur $\vec{0}m$ d'extrémité $m=(x_1, x_2)$. Sa densité d'énergie de déformation est supposée de la forme [16] $g_1(U(|y|, u, \nabla u))$ et son énergie totale est

$$(\mathcal{P}_1)\qquad J_1(u)=\int_\Omega g_1(U(|y|, u, \nabla u))\,dy+\lambda\int_\Omega h_1(|y|, u)\,dy$$

où U est une fonction assez spécifique que nous préciserons ultérieurement. De plus, signalons que ces fonctionnelles sont en général non convexes (cf. [9], [5], [1], [10], [6]).

Le modèle de mécanique exposé ci-dessus a fait l'objet d'une étude en [12] avec des hypothèses restrictives par rapport à notre travail. Notre méthode (cf. [4], [17]) est différente de celle mise en oeuvre en [12].

Nous aborderons également l'étude de fonctionnelles plus générales que les précédentes, de la forme

$$(\mathcal{P}_2)\qquad J_2(v)=\int_\Omega g_2(|x|, |\nabla v_1|, \cdots, |\nabla v_m|)+\lambda\int_\Omega h_2(|x|, v)\,dx$$

où $\nabla v=\partial v_j/\partial x_i$, $(1\leqq i\leqq n,\ 1\leqq j\leqq m)$, désigne la matrice gradient de v, et celles du type

$$(\mathcal{P}_3)\qquad J_3(v)=\int_\Omega g_3(Av)\,dx+\int_\Omega h(|x|, v)\,dx$$

où $Av=(A_1v_1, A_2v_2)$ est un opérateur différentiel uniformément elliptique.

Etant donnée une fonctionnelle $J(v)$ semi-continue inférieure faible séquentiellement (s.c.i.f.s.), la méthode directe du calcul des variations consiste à considérer une suite minimisante u_n convergeant vers u, et à passer à la limite dans $J(u_n)$. La s.c.i.f.s. permet alors d'avoir l'inégalité essentielle

$$\underline{\lim}\, J(u_n)\geqq J(u)$$

prouvant ainsi que u réalise le minimum de $J(\cdot)$.

La difficulté majeure que l'on rencontre dans l'étude des fonctionnelles introduites ci-dessus provient de l'absence de semi-continuité inférieure séquentielle faible qui rend inopérante la méthode directe du calcul des variations [9], [8].

Pour la résolution de ces problèmes, nous utilisons l'idée de base contenue dans [3]. La conclusion proviendra d'une analyse fine des conditions d'extrémalité pour obtenir des informations qualitatives sur les solutions des problèmes relaxés correspondants. Cette analyse fine est intimement liée à la structure de chaque problème. Dans cette étude, nous ne présenterons en détail que les points spécifiques aux problèmes abordés ici.

Nous traiterons également dans ce travail la question de l'unicité.

Enfin signalons que ce travail a été annoncé dans [18].

2. Hypothèses et notations du problème ($\mathcal{P}_1$). Pour simplifier l'exposé nous prendrons, d'une part des "fonctions modèles" U de type assez simple mais suffisamment expressives pour la méthode, et nous travaillerons, d'autre part, dans l'espace

fonctionnel $V = H^1(\Omega)$; ce qui nous amène à supposer les hypothèses de croissance convenables sur les fonctions g_1 et h_1. La première est une fonction de $\mathbb{R}$ dans $\mathbb{R}$ régulière, paire telle que

$$0 < \theta_0 \leqq t^{-1} \cdot \frac{dg_1^{**}}{dt}(t) \leqq \theta_1 \quad \forall t, \tag{1}$$

$$a_1 t^2 + b_1 \leqq g_1(t) \leqq a_2 t_2^2 + b_2 \quad \forall t, \tag{2}$$

où a_1 et a_2 sont des constantes positives et g_1^{**} désigne la convexifiée de g_1. La seconde est une fonction régulière de $[a, b] \times \mathbb{R}^3$ dans $\mathbb{R}$ satisfaisant

$$c_1|\eta|^2 + d_1 \leqq h_1(s, \eta) \leqq c_2|\eta|^2 + d_2 \quad \forall (s, \eta), \tag{3}$$

où c_2 est une constante positive et c_1 une constante convenablement choisie;

$$\frac{\partial h_1}{\partial \eta_i}(r, \eta) \leqq 0, \quad i = 2, 3, \quad \forall (r, \eta) \in]a, b[\times \mathbb{R}^3, \tag{4}$$

$$\frac{\partial h_1}{\partial \eta_1}(r, \eta) \equiv 0, \tag{5}$$

$$\frac{\partial h_1}{\partial \eta_3}(r, \eta) + \frac{\partial h_1}{\partial \eta_2}(r, \eta) < 0 \quad \forall (r, \eta) \in]a, b[\times \mathbb{R}^3. \tag{6}$$

La fonction U est supposée de la forme

$$U^2 = U^2(|x|, v, \nabla v) = |\nabla v_1|^2 + \beta^2(|x|, v_1) \cdot |\nabla v_2|^2 + |\nabla v_3|^2$$

où β est une fonction régulière qui vérifie

$$\beta(s, t) \geqq \beta_0 > 0, \tag{7}$$

$$\beta_1(s, t) = \frac{\partial \beta}{\partial s}(s, t) \geqq 0, \tag{8}$$

$$t \cdot \frac{\partial \beta}{\partial t}(s, t) \geqq 0, \; t^{-1} \cdot \frac{\partial \beta}{\partial t}(s, t) < c \quad \text{pour } t \text{ voisin de zéro.} \tag{9}$$

On note:

$$\Gamma_1 = \{x \in \mathbb{R}^2 / |x| = a\}, \qquad \Gamma_2 = \{x \in \mathbb{R}^2 / |x| = b\},$$

$$\Omega = \{x \in \mathbb{R}^2 / a < |x| < b\},$$

$$V_1 = (H_0^1(\Omega)^3 + \varphi, \tag{10}$$

où $\varphi = (\varphi_1, \varphi_2, \varphi_3)$ appartenant à $(H^1(\Omega))^3$ vérifie les conditions aux limites

$$\varphi / \Gamma_1 = \alpha = (\alpha_1, \alpha_2, \alpha_3), \qquad \varphi / \Gamma_2 = \gamma = (\gamma_1, \gamma_2, \gamma_3); \tag{11}$$

les α_i and γ_i sont des constantes supposées vérifier, pour fixer les idées,

$$\delta_i = \gamma_i - \alpha_i > 0, \qquad i = 2, 3, \tag{12}$$

$$\alpha_1 = 0, \qquad \gamma_1 > 0. \tag{13}$$

Nous voulons montrer que le problème

$(\mathscr{P}_1)$ Trouver une solution de $\inf\{J_1(v), v \in V_1\}$

admet au moins une solution, et que cette dernière possède une symétrie radiale. Cette propriété de la solution désirée nous amène à associer à $(\mathscr{P}_1)$ le problème relaxé, en dimension un d'espace:

(π_1) Trouver u dans $\tilde{V}_1$ solution de $\inf\{\tilde{J}_1(v), v \in \tilde{V}_1\}$,

l'espace $\tilde{V}_1$ et la fonctionnelle $\tilde{J}_1$ sont définis par

$$\tilde{V}_1 = \{v \in H^1((a,b))^3 / v_i(a) = \alpha_i,\ v_i(b) = \gamma_i\},$$

$$\tilde{J}_1(v) = \int_a^b r g_1^{**}(V(v))\, dr + \lambda \int_a^b r h_1(r, v)\, dr,$$

où l'on a noté

$$(V(v))^2 = \left(\frac{dv_1}{dr}\right)^2 + \beta^2(r, v_1) \cdot \left(\frac{dv_2}{dr}\right)^2 + \left(\frac{dv_3}{dr}\right)^2.$$

Pour simplifier la présentation nous noterons, si nécessaire, $v' = dv/dr$. Le problème (π_1) admet au moins une solution $u = u_\lambda$. L'idée est de montrer que u vérifie

$$g_1^{**}(V(u(r))) = g_1(V(u(r))) \quad \text{p.p. } r \in]a, b[.$$

Ce sera l'étape 1. L'étape 2 sera consacrée à la régularité de u. Enfin à l'étape 3, nous montrerons que la fonction $\bar{u}(x) = u(|x|)$ est solution de $(\mathscr{P}_1)$.

Etape 1. Cette étape comprend plusieurs propositions.

PROPOSITION 2.1. *On suppose les hypothèses* (1)-(3), (10), *et* (11). *La fonction u solution de* (π_1) *satisfait les relations d'extrémalités suivantes*:

$$r \cdot \frac{g_1^{**\prime}(V)}{V} \cdot u_1' = -p_1, \tag{14}$$

$$r \cdot \frac{g_1^{**\prime}(V)}{V} \beta \cdot \beta_2 \cdot (u_2')^2 + \lambda r \frac{\partial h_1}{\partial \eta_1}(r, u) = -p_1', \tag{15}$$

$$r \cdot \frac{g_1^{**\prime}(V)}{V} \cdot \beta^2 \cdot u_2' = -p_2, \tag{16}$$

$$r \cdot \frac{g_1^{**\prime}(V)}{V} u_3' = -p_3, \tag{17}$$

$$\lambda r \cdot \frac{\partial h_1}{\partial \eta_i}(r, u) = -p_i', \qquad i = 2, 3, \tag{18}$$

où $V = V(u)$, $\beta_2 = \partial\beta/\partial t(r, u_1)$, *et* $g_1^{**\prime}(t) = dg_1^{**}/dt$. *De plus les fonctions* p_i, u_i *appartiennent à* $W^{1,\infty}(a, b)$, $i = 1, 2, 3$.

Démonstration. La preuve de ce résultat s'obtient aisément à partir des équations d'Euler de π_1:

$$-\frac{d}{dr}(r\sigma \cdot u_1') + \lambda r \frac{\partial h_1}{\partial \eta_1}(r, u) + r\sigma \cdot \beta\beta_2 \cdot (u_2')^2 = 0,$$

$$-\frac{d}{dr}(r\sigma \cdot \beta^2 u_2') + \lambda r \frac{\partial h_1}{\partial \eta_2}(r, u) = 0,$$

$$-\frac{d}{dr}(r\sigma u_3') + \lambda r \frac{\partial h_1}{\partial \eta_3}(r, u) = 0.$$

Nous la laissons au lecteur.

Les p_i étant ainsi définies, considérons la fonction de $L^\infty(a, b)$

$$\nu_\lambda(r) = \nu(r) = \left(\frac{p_1}{r}\right)^2 + \left(\frac{p_2}{r\beta}\right)^2 + \left(\frac{p_3}{r}\right)^2;$$

nous avons le résultat suivant.

PROPOSITION 2.2. *Nous supposons* (1)-(8) *et* (10)-(12). *La fonction* $\nu(r)$ *appartient à* $W^{1,\infty}(a, b)$; *de plus, il existe* $\lambda_0 > 0$ *tel que pour tout* λ, $0 < \lambda < \lambda_0$, *on a* $\nu'_\lambda(r) < 0$, $u'_2(r) \geqq 0$, *et* $u'_3(r) \geqq 0$ *p.p.r dans* $]a, b[$.

Démonstration. $\nu(r)$ appartient à $W^{1,\infty}(a, b)$ car p_2^2, p_3^2 appartiennent à $C^1([a, b])$; et $1/\beta^2$ et p_1^2 appartiennent à $W^{1,\infty}(a, b)$. Ainsi, la dérivée au sens des distributions

$$\nu'(r) = \frac{2p_1p'_1}{r^2} + \frac{2p_2p'_2}{r^2\beta^2} + \frac{2p_3p'_3}{r^2} - \frac{2}{r}\left[\left(\frac{p_1}{r}\right)^2 + \left(\frac{p_2}{r\beta}\right)^2 + \left(\frac{p_3}{r}\right)^2\right] - \frac{2(\beta_1 + \beta_2 \cdot u'_1)}{\beta} \cdot \left(\frac{p_2}{r\beta}\right)^2$$

appartient à $L^\infty(a, b)$. Transformons quelque peu cette expression de $\nu'(r)$ à l'aide de (14), (15), et (16); nous obtenons

$$\nu'(r) = \frac{2}{r^2}\left[\frac{p_2p'_2}{\beta_2} + p_3 \cdot p'_3\right] - \frac{2}{r}\left[\left(\frac{p_1}{r}\right)^2 + \left(\frac{p_2}{r\beta}\right)^2 + \left(\frac{p_3}{r}\right)^2\right] - \frac{2\beta_1}{\beta} \cdot \left(\frac{p_2}{r\beta}\right)^2 \tag{19}$$

on voit bien que l'on aura $\nu'(r) < 0$ si l'on montre par exemple que

$$p_3 \cdot p'_3 + \frac{p_2p'_2}{\beta^2} < 0. \tag{20}$$

Remarque 2.1. Les solutions (u_i, p_i), $i = 1, 2, 3$ du système (14)-(18) dépendent du paramètre λ, i.e., que l'on a en fait

$$u_i = u_i^\lambda, \quad p_i = p_i^\lambda, \quad i = 1, 2, 3.$$

On montre aisément les estimations suivantes:

$$\begin{aligned} &\|u_i^\lambda\|'_{L^\infty(a,b)} \leqq c, && i = 1, 2, 3, \\ &\|u_i^\lambda\|_{H^1(a,b)} \leqq c, && i = 1, 2, 3, \\ &\|p_i^\lambda\|_{H^1(a,b)} \leqq c, && i = 1, 2, 3, \end{aligned} \tag{21}$$

où $c = c(g_1, h_1, a, b, \delta_i)$ désigne diverses constantes indépendantes de λ.

Montrons (20). Par une intégration par parties à partir de (16) et (17), on montre les inégalités suivantes:

$$\begin{aligned} 0 < \delta_2 &= u_2(b) - u_2(a) \\ &= -p_2(b)\Sigma_2(b) - \lambda \int_a^b r \frac{\partial h_1}{\partial \eta_2}(r, u) \cdot \Sigma_2 \, dr, \\ 0 < \delta_3 &= u_3(b) - u_3(a) \\ &= -p_3(b)\Sigma_3(b) - \lambda \int_a^b r \frac{\partial h_1}{\partial \eta_3}(r, u) \cdot \Sigma_3 \, dr \end{aligned}$$

où

$$\Sigma_2(r) = \int_a^r \frac{V \cdot ds}{\beta^2 \cdot s \cdot g_1^{**\prime}(V)}, \qquad \Sigma_3(r) = \int_a^r \frac{V \cdot ds}{s \cdot g_1^{**\prime}(V)}.$$

Et ainsi, grâce aux estimations (21), il existe $\lambda_0 > 0$ tel que pour tout λ dans $]0, \lambda_0[$ on a

$$-p_i(b)\Sigma_i(b) > 0, \qquad i = 2, 3,$$

i.e.,

$$p_i(b) < 0, \qquad i = 2, 3;$$

ceci entraîne que les fonctions p_2 et p_3 sont strictement négatives puisque nous avons supposé l'hypothèse (4) à savoir

$$\frac{\partial h_1}{\partial \eta_i}(r, \eta) \leqq 0, \qquad i = 2, 3.$$

Et enfin l'hypothèse (6) entraîne que (20) est vérifié, i.e., $\nu'(r) < 0$. Ainsi s'achève la preuve de la proposition 2.2. □

Remarque 2.2. Une question naturelle se pose: a-t-on une propriété de monotonie pour u_1?

La réponse à la question ci-dessus nous sera donnée par la proposition 2.4.

PROPOSITION 2.3. *Sous les hypothèses* (1) *à* (8) *et* (10) *à* (12), *toute solution u du problème* (π_1) *satisfait*

$$g_1^{**}(V(u(r))) = g_1(V(u(r))) \quad p.p.\ r \in]a, b[.$$

Démonstration. Il suffit de montrer que l'ensemble

$$E = \{r \in]a, b[/ g_1^{**}(V(u(r))) < g_1(V(u(r)))\}$$

est de mesure nulle. Cela se montre, comme dans [17], à l'aide de la propriété suivante de la fonction ν conséquence de la proposition 2.2:

$$|\{r \in [a, b] / \nu(r) = t\}| = 0 \quad \forall t.$$

Remarque 2.3. La fonction ν étant sans palier, il est aisé de voir que l'on a de plus le résultat suivant: s'il existe une partie affine commune aux graphes de g_1 et g_1^{**}, soit par exemple

$$\Delta = \{(t, g_1(t)), t \in \tilde{K}\} \subset \mathbb{R}^2$$

où

$$\tilde{K} = \{t \in \mathbb{R} / g_1'(t) = g_1^{**\prime}(t) = c = \text{constante}\}$$

alors l'ensemble $\{r \in]a, b[/ V(u(r)) \in K\}$ est de mesure nulle. Ainsi "n'interviennent" que les parties strictement convexes du graphes de g_1^{**}.

Cette information nous sera utile pour obtenir un résultat d'unicité (cf. le théorème 2.2).

Revenons maintenant à la monotonie de u_1. Pour cela nous aurons besoin des hypothèses supplémentaires (9) et (13).

PROPOSITION 2.4. *Si l'on suppose que* (9) *et* (13) *ont lieu, alors on a*

$$p_1(r) \leqq 0 \quad \forall r,$$

$$u_1'(r) \geqq 0 \quad p.p.\ r,$$

$$0 \leqq u_1(r) \leqq \gamma_1 \quad \forall r.$$

Démonstration. La fonction u_1 vérifie l'équation

$$-\frac{d}{dr}\left(r\sigma \frac{du_1}{dr}\right) + r\sigma\beta\left(\frac{\beta_2}{u_1}\right)(u_2')^2 \cdot u_1 = 0,$$

$$u_1(a) = 0, \qquad u_1(b) = \gamma_1 > 0,$$

avec $r\sigma \geqq a$, $\theta_0 > 0$, $(u_2')^2 r\sigma\beta \cdot \beta_2 / u_1 \geqq 0$. Le principe du maximum entraîne que l'on a $u_1(r) > 0$ sur $]a, b[$; d'autre part on a par (15) $p_1'(r) \leqq 0$ p.p. r, i.e., que p_1 est décroissant. Et pour montrer que p_1 est négatif nous raisonnerons par l'absurde. Supposons qu'il existe $r_0 \in]a, b]$ tel que $p_1(r) > 0$ $\forall r \in]a, r_0[$; (14) entraîne que l'on a $u_1'(r) \leqq 0$ p.p. $r \in]a, r_0[$ i.e., $u_1(r) \leqq 0$ pour tout r dans $]a, r_0[$. Ceci contredit le principe du maximum. La conclusion suit. □

Remarque 2.4. A travers la preuve de ce résultat il est aisé de voir que si $u_1(a) = \alpha_1 > 0$, il peut exister $r_0 \in]a, b[$ tel que l'on ait

$$u_1'(r) \leqq 0 \quad \text{p.p. sur }]a, r_0[,$$

$$u_1'(r) \geqq 0 \quad \text{p.p. sur }]r_0, b[,$$

$$u_1(r) \geqq 0 \quad \text{sur }]a, b[.$$

Remarque 2.5. Si l'hypothèse (5) n'a plus lieu et si par exemple

$$h_1(r, \eta) = \lambda h(r, \eta_2, \eta_3) + \gamma k(r, \eta_1), \quad 0 < \lambda \leqq \lambda_0, \quad 0 < \gamma \leqq \gamma_0$$

étant deux paramètres réels indépendants, alors le résultat de la proposition 2.2 subsiste pour λ_0 et γ_0 suffisamment petits. En revanche, nous ne pouvons obtenir un résultat de monotonie sur u_1, propriété dont on aura besoin à la § 4 où est traité un exemple issu de l'élasticité.

Etape 2. Régularité de u solution de (π_1). D'après la proposition 2.2, la fonction $\nu(r)$ est continue strictement décroissante. Introduisons, suivant [17], les notations suivantes qui nous seront utiles pour le résultat d'unicité:

$$E(\Lambda) = \left\{ \gamma \in \mathbb{R}^+ \Big/ \frac{dg_1^{**}}{dt}(\gamma) = \Lambda \right\}, \qquad \Lambda \in \mathbb{R}^+,$$

$$a(\Lambda) = \inf\{\gamma, \gamma \in E(\Lambda)\}, \qquad b(\Lambda) = \sup\{\gamma, \gamma \in E(\Lambda)\},$$

$$S = \{\Lambda \in \mathbb{R}^+ / a(\Lambda) < b(\Lambda)\}, \qquad \psi(t) = \frac{dg_1^{**}}{dt}(t).$$

Soit $\Lambda_0 < \Lambda_1 < \Lambda_2 < \cdots < \Lambda_n \cdots$ les éléments de S. La fonction continue

$$\phi(r) = \sqrt{\nu(r)} = \frac{dg_1^{**}}{dt}(V(u(r)))$$

étant strictement décroissante, il existe une suite de nombres réels

$$\{s_k\}_{k=0}^{k=n_1} \subset [a, b]$$

satisfaisant

$$\phi(s_k) = \Lambda_k, \qquad k = 0, 1, \cdots, n_1.$$

D'après la remarque 2.3 la fonction

$$(\psi_{/]b(\Lambda_k), a(\Lambda_{k+1})[})(t)$$

est inversible puisque g_1^{**} est strictement convexe sur $]b(\Lambda_k), a(\Lambda_{k+1})[$. Ainsi:

$$V(u(r)) = [\psi_{/]b(\Lambda_k), a(\Lambda_{k+1})[}]^{-1}(\phi(r))$$

pour tout $s \in]s_k, s_{k+1}[$, i.e., la fonction $V(u(r))$ est continue sur $]s_k, s_{k+1}[$ tout $k = 0, 1, \cdots, n_1$. Ceci prouve à l'aide de (1) et de la proposition 2.2 que $u(r)$ et $\phi(r)$ sont

de classe C^1 sur $]s_k, s_{k+1}[$; donc $V(u(r))$ est C^1 sur $]s_k, s_{k+1}[$; et ainsi de suite on montre que u est C^∞ par morceaux si g_1 et h_1 sont C^∞.

Ainsi, nous venons de démontrer le résultat suivant:

PROPOSITION 2.5. *La fonction u solution de* (π_1) *est* C^∞ *par morceaux si* g_1 *et* h_1 *sont* C^∞.

Etape 3. $\bar{u}(x) = u(|x|)$ est solution de $(\mathcal{P}_1)$.

Comme conséquence des 2 étapes précédentes, on obtient le résultat suivant.

THÉORÈME 2.1. *Sous les hypothèses* (1) *à* (8) *et* (10) *à* (12), *il existe* λ_0 *tel que le problème* $(\mathcal{P}_1)$ *admet, pour tout* λ *dans* $]0, \lambda_0[$, *une solution* $\bar{u}$ *possédant les propriétés suivantes*:

(1) $\bar{u}(x)$ *est radiale*: $\bar{u}(x) = \bar{u}(|x|)$,
(2) $|\nabla \bar{u}(x)| \leqq C$ *p.p.* $x \in \Omega$
(3) $\bar{u}$ *est* C^∞ *par morceaux si* g_1 *et* h_1 *sont* C^∞,
(4) $\bar{u}_2$ *et* $\bar{u}_3$ *sont croissantes le long des rayons.*

De plus si (9) *et* (13) *ont lieu* $\bar{u}_1$ *est croissante le long des rayons.*

Démonstration. Il suffit de voir que $\bar{u}(x) = u(|x|)$ convient. □

Abordons maintenant l'unicité. Rappelons que ce point n'a pas été abordé dans [4] et [17]. Nous avons le résultat suivant.

THÉORÈME 2.2. *On suppose* $\beta(r, t) = \beta(r)$, h *convexe, et* (1)-(8) *et* (12). *Alors* (π_1) *admet une solution unique.*

Démonstration. La preuve de ce résultat nécessite une étape préliminaire: le lemme 2.1. Soient $u \not\equiv \omega$ deux solutions de (π_1). Elles vérifient

$$g_1^{**}(V(u(r))) = g_1(V(u(r))) \quad \text{p.p. } r \in]a, b[,$$

$$g_1^{**}(V(\omega(r))) = g_1(V(\omega(r))) \quad \text{p.p. } r \in]a, b[;$$

pour tout θ dans $[0, 1]$, la fonction $u_\theta = \theta u + (1-\theta)\omega$ solution de (π_1) vérifie la même relation:

$$g_1^{**}(V(u_\theta(r))) = g_1(V(u_\theta(r))) \quad \text{p.p. } r \in]a, b[\quad \forall \theta \in [0, 1]. \tag{22}$$

D'après le résultat de régularité il existe une partition de $]a, b[$ en intervalles $J_k =]s_k s_{k+1}[$, $k = 1, 2, \cdots, n_1$ tels que: $u'(r)$ et $V(u(r))$ soient continues sur J_k pour tout k. Quitte à redécouper les J_k en sous-intervalles, on peut toujours supposer que $\omega'(r)$ et $V(\omega(r))$ sont continues sur J_k pour tout k. Posons alors

$$\mathcal{U}(r, \theta) = V(u_\theta(r))$$

$$= [(u'_{1\theta})^2 + \beta^2(r) \cdot (u'_{2\theta})^2 + (u'_{3\theta})^2]^{1/2}$$

où $u_\theta(r) = \theta u(r) + (1-\theta)\omega(r)$. Il est clair que $\mathcal{U}(r, \theta)$ est continue sur $J_k \times]0, 1[$ à valeurs dans

$$F = \{t \in \mathbb{R}^+ / g_1^{**}(t) = g_1(t)\}.$$

Nous avons le résultat suivant.

LEMME 2.1. *Nous avons* $J_k \cap V(u)^{-1}(I) = J_k \cap V(\omega)^{-1}(I)$ *pour tout* k *et toute composante connexe* I *de* F.

Démonstration. Il suffit de montrer que l'on a

$$[V(u)^{-1}(I_1) \cap J_k] \cap [V(\omega)^{-1}(I_2) \cap J_k] = \varnothing$$

pour tout k et toutes composantes connexes de F, I_1, et I_2 telles que $I_1 \neq I_2$. Nous raisonnerons par l'absurde.

Supposons qu'il existe $k_0 \in [1, \cdots, n_1]$ et deux composantes connexes I_1 et I_2 $(I_1 \neq I_2)$, de F tels que l'ensemble

$$\Omega = V(u)^{-1}(I_1) \cap V(\omega)^{-1}(I_2) \cap J_{k_0}$$

soit non vide; soit $r_0 \in \Omega$; nous avons

$$V(u(r_0)) = \mathcal{U}(r_0, 1) \in I_1, \qquad V(\omega(r_0)) = \mathcal{U}(r_0, 0) \in I_2,$$

et puisque $\mathcal{U}(J_{k_0} \times]0, 1[)$ est connexe, il existe $\theta_0 \in]0, 1[$, δ_1 et δ_2 tels que

$$\mathcal{U}(r_0, \theta_0) \in]\delta_1, \delta_2[\subseteq K$$

où K est le complémentaire de F. Et par continuité il existe un voisinage $\mathcal{V}(r_0)$ tel que

$$\mathcal{U}(r, \theta_0) \in]\delta_1, \delta_2[\quad \forall r \in \mathcal{V}(r_0),$$

c'est à dire

$$g_1(V(u_{\theta_0}(r))) > g_1^{**}(V(u_{\theta_0}(r))) \quad \text{p.p. } r \in \mathcal{V}(r_0)$$

contredisant ainsi (22). Ainsi le lemme 2.1 a lieu. □

Démonstration du théorème 2.2. Si $u \not\equiv \omega$, d'après la remarque 2.3 et le lemme 2.1, il existe $k_0 \in [1, \cdots, n_1]$ tel que

$$\int_{J_{k_0}} g_1^{**}(V(u_{1/2}(r)))\, dr < \frac{1}{2} \int_{J_{k_0}} g_1^{**}(V(u(r)))\, dr + \frac{1}{2} \int_{J_{k_0}} g_1^{**}(V(\omega(r)))\, dr;$$

cette inégalité entraîne la contradiction

$$J_1(u_{1/2}) < \operatorname{Inf}\{J_1(v), v \in V_1\}.$$

Remarque 2.6. Le théorème 2.2 entraîne que la solution radiale de $(\mathcal{P}_1)$ est unique.

3. Hypothèses et notation du problème (P_2). Soit la fonction g_2 de $[a, b] \times \mathbb{R}^2$ dans $\mathbb{R}$, régulière telle que

$$a_1|t|^2 + b_1(r) \leqq g_2(r, t) \leqq a_2|t|^2 + b_2(r) \quad \forall (r, t) \tag{1}$$

où les a_i désignent des constantes positives et les b_i des fonctions $L^\infty(a, b)$. Désignons par $g_2^{**}(r, \cdot)$ la convexifiée de $g_2(r, \cdot)$ par rapport à t, à r fixé, et posons:

$$\mathcal{K} = \{(r, t) \in [a, b] \times \mathbb{R}^2 / g^{**}(r, t) < g_2(r, t)\}. \tag{2}$$

On suppose que, d'une part les composantes connexes $\mathcal{K}_i$ de $\mathcal{K}$, indexées par I, sont au plus dénombrables:

$$\mathcal{K} = \bigcup_{i \in I} \mathcal{K}_i; \tag{3}$$

d'autre part, pour tout i dans I il existe une fonction

$$k_i : (r, t) \in]a, b[\times \mathbb{R}^2 \to k_i(r, t) \in \mathbb{R}$$

affine en t telle que

$$g_2^{**}(r, t) = k_i(r, t) = \sigma_i(r) \cdot t + \xi_i(r) \quad \forall (r, t) \in \mathcal{K}_i \tag{4}$$

où $\sigma_i = (\sigma_i^1, \sigma_i^2)$ sont deux fonctions de $W^{1,\infty}(a, b)$ satisfaisant:

$$\sum_{j=1}^{2} \sigma_i^j(r) \cdot \sigma_i^{j\prime}(r) \geqq 0 \quad \text{p.p. } r \in]a, b[. \tag{5}$$

Cette hypothèse (4), (5) jouera un rôle essentiel dans l'existence. On suppose aussi que g_2 vérifie

$$
\begin{aligned}
&0 < \theta_0 \leqq t_i^{-1} \cdot \frac{\partial g_2^{**}}{\partial t_i}(r, t) \leqq \theta_i(r, t_j) \quad \forall t = (t_1, t_2) \quad \forall i \neq j, \\
&0 \leqq \theta_i(r, t_j) \leqq \gamma_i
\end{aligned} \tag{6}
$$

où θ_0 et γ_i sont des constantes positives:

$$
\begin{aligned}
&g_2(r, t_1, t_2) = g_2(r, \varepsilon_1 t_1, \varepsilon_2 t_2) \quad \forall (r, t), \\
&\varepsilon_1 = \pm 1, \qquad \varepsilon_2 = \pm 1.
\end{aligned} \tag{7}
$$

On se donne une fonction h_2 de $[a, b] \times \mathbb{R}^2$ dans $\mathbb{R}$, régulière telle que

$$
c_1 |\eta|^2 + d_1 \leqq h_2(r, \eta) \leqq c_2 |\eta|^2 + d_2 \quad \forall (r, \eta), \tag{8}
$$

où c_2 est une constante positive et c_1 une constante réelle convenable:

$$
\frac{\partial h_2}{\partial \eta_i}(r, \eta) < 0 \quad \forall (r, \eta), \quad i = 1, 2. \tag{9}
$$

Nous nous proposons alors de résoudre le problème suivant:

$(\mathscr{P}_2)$ Trouver $\bar{u}$ dans V_2 solution de $\inf \{J_2(v), v \in V_2\}$

où

$$
J_2(v) = \int_\Omega g_2(|x|, |\nabla v_1|, |\nabla v_1|)\, dx + \lambda \int_\Omega h_2(|x|, v)\, dx,
$$

$$
V_2 = \{v \in (H^1(\Omega))^2 / v_{i/\Gamma_1} = d_i, v_{i/\Gamma_2} = \beta_i\};
$$

α_i et β_i sont des constantes données, qu'on suppose satisfaisant, pour simplifier, l'inégalité

$$
\alpha_i < \beta_i, \qquad i = 1, 2. \tag{10}
$$

L'idée utilisée pour résoudre $(\mathscr{P}_1)$ reste valable: on part du problème variationnel, en dimension un d'espace, qui est naturellement associé à $(\mathscr{P}_2)$:

(π_2) Trouver $u \in \tilde{V}_2$ solution de $\inf \{\tilde{J}_2(v), v \in \tilde{V}_2\}$

avec

$$
\tilde{J}_2(v) = \int_a^b r g_2^{**}(r, v_1', v_2')\, dr + \lambda \int_a^b r h_2(r, v)\, dr,
$$

$$
\tilde{V}_2 = \{v \in (H^1(a, b))^2 / v_i(a) = \alpha_i, v_i(b) = \beta_i\}.
$$

Ce problème relaxé (π_2) admet au moins une solution $u = u^\lambda$ satisfaisant les relations d'extrémalités

$$
\begin{aligned}
&r \frac{\partial g_2^{**}}{\partial t_i}(r, u') = -p_i \in H^1(a, b), \qquad i = 1, 2, \\
&r\lambda \frac{\partial h_2}{\partial \eta_i}(r, u) = -p_i', \qquad i = 1, 2.
\end{aligned} \tag{11}
$$

Nous allons montrer (comme précédemment) à l'aide des hypothèses (3)-(5) et (9) que pour tout $i \in I$ l'ensemble

$$E_i = \{r \in]a, b[/(r, u'(r)) \in \mathcal{K}_i\}$$

est de mesure nulle, i.e.,

$$g_2^{**}(r, u'(r)) = g_2(r, u'(r)) \quad \text{p.p. } r \in]a, b[.$$

PROPOSITION 3.1. *La solution $u = u^\lambda$ de (π_2) satisfait*

$$u_i' = (u_i^\lambda)' \geqq 0.$$

Démonstration. Pour λ dans $]0, 1]$ on montre les estimations suivantes:

$$\|u_i\|_{H^1} = \|u_i^\lambda\|_{H^1(a,b)} \leqq c, \qquad \|p_i\|_{H^1} = \|p_i^\lambda\|_{H^1(a,b)} \leqq c$$

où c est une constante indépendante de λ. A l'aide d'une intégration par parties à partir de (11) on obtient les relations

$$\begin{aligned} 0 < \delta_i &= u_i(b) - u_i(a) \\ &= -p_i(b)\Sigma_i(b) - \lambda \int_a^b r \frac{\partial h_2}{\partial \eta_i}(r, u) \cdot \Sigma_i(r)\, dr \end{aligned}$$

où Σ_i a pour expression

$$\Sigma_i^\lambda(r) = \Sigma_i(r) = \int_a^r u_i'(s) \cdot \frac{1}{s} \left(\frac{\partial g_2^{**}}{\partial t_i}(s, u') \right)^{-1} ds, \qquad i = 1, 2.$$

On peut alors affirmer l'existence d'un λ_0 dans $]0, 1]$ tel que

$$-p_i^\lambda(b)\Sigma_i^\lambda(b) > 0 \quad \forall \lambda \in]0\lambda_0].$$

Cette inégalité entraîne $p_i(b) < 0$ puisque $\Sigma_i(b)$ est strictement positif à l'aide de (6); ainsi on a

$$p_i(r) < 0 \quad \forall r \in [a, b], \quad i = 1, 2, \tag{12}$$

en vertu de l'hypothèse (9). Par conséquent on peut dire, utilisant l'hypothèse (6), que

$$0 \leqq r \cdot u_i' \cdot \frac{\partial g_2^{**}}{\partial t_i}(r, u') = -p_i \cdot u_i',$$

i.e.,

$$u_i'(r) \geqq 0 \quad \text{p.p. } r \in [a, b], \quad i = 1, 2.$$

Ceci termine la preuve de la proposition. □

PROPOSITION 3.2. *La fonction u vérifie*

$$g_2^{**}(r, u'(r)) = g_2(r, u'(r)) \quad \text{p.p. } r \in]a, b[. \tag{13}$$

Démonstration. Considérons la fonction $\nu(r) = \sum_{i=1}^2 (p_i(r)/r)^2$; à l'aide de (12) et (9) on montre que $\nu'(r) < 0$ p.p. $r \in]a, b[$. Comme pour tout i and I la fonction $\mu_i(r) = \sum_{j=1}^2 (\sigma_i^j)^2$ est croissante, il s'ensuit que l'ensemble $\{r \in]a, b[/\mu_i(r) = \nu(r)\}$ est de mesure nulle; ce qui entraîne que pour tout i dans I l'ensemble $E_i = \{r \in]a, b[/(r, u'(r)) \in \mathcal{K}_i\}$ est de mesure nulle. Ainsi (13) a lieu.

Enfin nous pouvons montrer facilement que u_1 appartient à $W^{1,\infty}(a, b)$. Une conséquence immédiate de ces résultats est le théorème suivant.

THÉORÈME 3.1. *On suppose les hypothèses* (1) *à* (10). *Alors il existe* $\lambda_0 > 0$ *tel que pour tout* λ *dans* $]0, \lambda_0[$ *le problème* $(\mathscr{P}_2)$ *admet au moins une solution* $\bar{u}$. *Cette solution possède les propriétés suivantes*:

(1) $\bar{u}(x)$ *est radiale, i.e.,* $\bar{u}(x) = \bar{u}(|x|)$;

(2) $|\nabla \bar{u}(x)| \leqq c$ *p.p.* $x \in \Omega$;

(3) $\bar{u}_i$ *est monotone le long de chaque rayon.*

Remarque 3.1. Il est possible d'envisager des fonctions g_2 dépendant de r, u, u'; mais cela nécessite des hypothèses supplémentaires qu'il serait fastidieux de décrire ici.

4. Application. Nous considérons un tube T homogène isotrope de section la couronne $C = [0, 2\pi[\times]a, b[$. Nous supposons qu'avant déformation ce corps élastique a ses génératrices parallèles à l'axe $0\vec{z}$:

$$T = C \times [-L, +L] = \{(r, \theta, z) \in]a, b[\times [0, 2\pi] \times [-L, L]\}$$

où nous notons

$$x = (x_1, x_2), \quad r = \sqrt{x_1^2 + x_2^2}, \quad x_1 = r \cos\theta, \quad x_2 = r \sin\theta.$$

Nous nous intéressons aux déformations de la forme

$$(1) \qquad (x, z) = (r, \theta, z) \to (u(r), \theta + v(r), z + w(r)) = (r', \theta', z').$$

D'après [16], [2], les trois invariants principaux du tenseur de Cauchy-Green ont pour expressions:

$$(2) \qquad \begin{aligned} I_1 &= (u')^2 + (uv')^2 + (w')^2 + \left(\frac{u}{r}\right)^2 + 1, \\ I_2 &= (u')^2 + (uv')^2 + \left(\frac{uu'}{r}\right)^2 + \left(\frac{uw'}{r}\right)^2 + \left(\frac{u}{r}\right)^2, \\ I_3 &= \frac{uu'}{r}. \end{aligned}$$

Dans la situation la plus générale, la fonction densité d'énergie de déformation a pour expression (cf. la remarque 3.1)

$$W(I_1, I_2, I_3) = g(r, u, u', v', w').$$

Et dans ce cas la réponse au problème n'est pas simple, notamment à cause de la condition d'invertibilité

$$(3) \qquad \frac{u(r) \cdot u'(r)}{r} > 0.$$

Aussi, pour illustrer simplement ce qui précède, nous allons examiner quelques situations particulières.

Exemple 1. Le cas incompressible. La déformation est de la forme

$$\begin{aligned} &(r, \theta, z) \to (r, \theta + v(r), z + w(r)) \\ &I = I_1 = I_2 = r^2 (v')^2 + (w')^2 + 3, \qquad I_3 = 1, \\ &W(I) = g(\sqrt{I - 3}). \end{aligned}$$

Nous représentons par $U = (v, w)$ le déplacement et par $\partial T = \partial_1 T \cup \partial_2 T$ le bord de T où

$$(4) \qquad \begin{aligned} \partial_1 T &= \{(x, z) / z = \pm L\}, \\ \partial_2 T &= \{(x, z) / |x| = a \text{ ou } |x| = b\}. \end{aligned}$$

Nous supposons que le déplacement est fixé sur $\partial_2 T$:

$$(v, w) = (v_0, w_0) \quad \text{sur } \partial_2 T, \tag{5}$$

et que la traction t est nulle sur $\partial_1 T$:

$$t(x, \pm L) = 0. \tag{6}$$

Alors le problème de l'équilibre élastostatique de T s'écrit: trouver $\bar{u} = (\bar{v}, \bar{w})$ réalisant dans l'espace $V = (H_0^1(C))^2 + (v_0, w_0)$, par exemple, le minimum de la fonctionnelle

$$J(v, w) = \int_C g(\sqrt{I-3})\, de + \lambda \int_C h(|x|, v, w)\, dx \tag{7}$$

où h représente la prise en compte des forces volumiques. Le théorème 2.1 s'applique.

Exemple 2. Le cas faiblement compressible. Nous supposons que la loi de comportement du matériau est telle que son énergie soit de la forme $g(I)$, où

$$\begin{aligned} I^2 &= I_1 - 2I_3 - 1 \\ &= r^2\left[\left(\frac{u}{r}\right)'\right]^2 + r^2\left(\frac{u}{r}\right)^2 (v')^2 + (w')^2. \end{aligned}$$

Remarque 4.1. Des situations particulières du même type ont été envisagées par divers auteurs: dans [13] et [14] par exemple, on considère des densités dépendant uniquement du premier invariant I_1.

Transformons quelque peu l'expression de I en posant

$$\begin{aligned} u(r) &= r + r \cdot y(r) = r(1 + y(r)), \\ I^2 &= r^2(y')^2 + r^2(1+y)^2(v')^2 + (w')^2. \end{aligned}$$

On s'impose les conditions aux limites suivantes:

$$\begin{aligned} &y(a) = 0, \quad \text{i.e., } u(a) = a, \\ &y(b) = y_1 > 0, \quad \text{i.e., } u(b) = b(1+y_1), \\ &v(a) = v_0, \qquad v(b) = v_1, \\ &w(a) = w_0, \qquad w(b) = w_1; \end{aligned} \tag{8}$$

la donnée y_1 sera prise assez petite. Ceci sera justifié plus loin. Il est difficile de contrôler une minoration du terme positif $1 + y(r)$. Aussi nous perturbons la fonctionnelle $J(\cdot)$ par

$$\tilde{J}(y, v, w) = \int_a^b g(\tilde{I}) r\, dr + \lambda \int_a^b h(r, y, v, w) r\, dr \tag{9}$$

où

$$\tilde{I} = r^2(y')^2 + r^2(1 + \sigma_\delta(y))^2 (v')^2 + (w')^2,$$

la fonction $\sigma_\delta(y)$ étant la régularisation classique de $|y|$:

$$\sigma_\delta(y) = \int_0^y \psi_\delta(\theta)\, d\theta,$$

$$\psi_\delta(\theta) = \begin{cases} 1 & \text{si } \theta \geqq \delta, \\ \dfrac{1}{\delta} \cdot \theta & \text{si } \delta > \theta > -\delta, \\ -1 & \text{si } \theta \leqq -\delta, \end{cases}$$

δ étant un paramètre réel, positif assez petit. Il est aisé de voir que l'on peut appliquer les résultats du paragraphe 2 en s'imposant les hypothèses adéquates du théorème 2.1;

et en particulier on obtient que la solution $(\bar{y}, \bar{v}, \bar{w})$ du problème (9), $\inf\{\tilde{J}(y, v, w)/(y, v, w)\}$, vérifie:

$$0 < \bar{y}(r) \leqq y_1 \quad \forall r \in]a, b[,$$

$$\bar{y}'(r) \geqq 0 \quad \text{p.p. } r, \qquad \|\bar{y}'\|_{L^\infty(a,b)} \leqq C_0,$$

ce qui donne bien la condition

$$\frac{uu'}{r} = (1+\bar{y})(1+\bar{y}+r\bar{y}') > 0 \quad \text{p.p. } r.$$

Le modèle de fonction énergie que nous avons considéré est valable pour les matériaux dont le comportement est voisin de celui des matériaux incompressibles (i.e., faiblement compressibles). Il est utile de savoir, lorsque y_1 tend vers zéro, si notre problème (8) "tend" vers le problème incompressible (6), et dans quel sens.

Nous avons le résultat suivant.

PROPOSITION 4.1. *On se donne une suite $y_{1,n} > 0$ tendant vers zéro; et on considère (y_n, v_n, w_n) une solution de* (8). *Alors on a*:

(i) $y_n \to 0$ *uniformément*;

(ii) $\|y_n'\|_{L^\infty(a,b)} \to 0$, $v_n \to \bar{v}$, $w_n \to \bar{w}$ *dans* $H^1(a, b)$ *faible*;

(iii) $(\bar{v}, \bar{w})$ *est solution du problème incompressible* (6);

(iv) $u_n u_n'/r \to 1$ *dans* $L^\infty(a, b)$.

La preuve s'obtient en établissant les estimations adéquates en suivant la même démarche que dans les paragraphes précédents.

Remarque 4.2. La perturbation suivante:

$$\tilde{I} = r^2(y')^2 + r^2(1+|y|)^2(v')^2 + (w')^2 \tag{10}$$

permet également d'obtenir que $\bar{y}$ est positif; cependant elle n'est pas satisfaisante pour notre méthode car $\tilde{I}$ n'est pas différentiable par rapport à y.

Remarque 4.3. Il serait intéressant de savoir si l'on peut passer à la limite quand δ tend vers zéro et obtenir la résolution de notre problème avec (10).

5. Fonctionnelles d'opérateurs elliptiques. Soit Ω un ouvert borné régulier de $\mathbb{R}^n$; on se donne, par exemple, deux opérateurs A_1 et A_2 uniformément elliptiques d'ordre $2m$, à coefficients réguliers, et deux fonctions régulières g_3 de $\Omega \times \mathbb{R}^2$ dans $\mathbb{R}$ et h_3 de $\Omega \times \mathbb{R}^2$ dans $\mathbb{R}$ satisfaisant les hypothèses classiques de croissance suivantes:

$$a_1|t|^p + b_1 \leqq g_3(x, t) \leqq a_2|t|^p + b_2, \qquad 1 < p < +\infty, \tag{1}$$

$$c_1|\eta|^p + d_1 \leqq h_3(x, \eta) \leqq c_2|\eta|^p + d_2; \tag{2}$$

les constantes a_1, a_2, c_2 sont positives; c_1 est un réel convenablement choisi pour assurer la coercitivité du problème considéré. On se donne également les deux opérateurs suivants:

$$A_i = \sum_{\substack{|\alpha|=m \\ |\beta|=m}} (-1)^{|\alpha|} D^\alpha (a_{\alpha\beta}^i D^\beta), \quad a_{\alpha\beta}^i = a_{\beta\alpha}^i, \quad i = 1, 2$$

dont les coefficients sont supposés réguliers. Et on pose

$$K = \{(x, t) \in \Omega \times \mathbb{R}^2 / g_3^{**}(x, t) < g_3(x, t)\}. \tag{3}$$

Nous faisons les hypothèses essentielles suivantes: il existe une famille I, au plus dénombrable de fonctions $\alpha_i = (\alpha_i^1, \alpha_i^2)$ dans $(W^{2m,p}(\Omega))^2$, $\beta_i \in L^\infty(\Omega)$, d'ensembles

$K_i \subset \Omega \times \mathbb{R}^2$ et une constante c tels que

$$K = \bigcup_{i \in I} K_i \quad \forall i \in I,$$

(4)
$$g_{3/K_i}^{**}(x, t) = \alpha_i(x) \cdot t + \beta_i(x) \quad \forall (x, t) \in K_i,$$

(5)
$$\left| A_1\alpha_1(x) + \frac{\partial h_3}{\partial \eta_1}(x, \eta) + c\left(A_2\alpha_2(x) + \frac{\partial h_3}{\partial \eta_2}(x, \eta) \right) \right| > 0 \quad \forall \eta \in \mathbb{R}^2 \quad \text{p.p. } x \in \Omega.$$

Remarque. Une condition du type (5) est donnée dans [4] dans le cas scalaire, régulier et dans [5] dans le cas scalaire non régulier.

Enfin notons par V_3 l'espace $(W^{2m,p}(\Omega) \cap W_0^{m,p}(\Omega))^2 + \varphi$, où φ est une fonction donnée dans $(W^{2m,p}(\Omega))^2$; et soit à minimiser la fonctionnelle

$$J_3(v) = \int_\Omega g_3(x, A_1v_1, A_2v_2)\, dx + \int_\Omega h_3(x, v)\, dx$$

sur l'espace V_3. Alors nous avons le résultat suivant.

THÉORÈME 4.1. *Le problème* $(\mathcal{P}_3)$: $\inf\{J_3(v)/v \in V_3\}$ *admet au moins une solution* $u = (u_1, u_2) \in V_3$.

Idée de la preuve. Le problème relaxé

$$(\mathcal{PR}) \qquad \inf\left\{ \int_\Omega g_3^{**}(x, Av)\, dx + \int_\Omega h_3(x, v)\, dx / v \in V_3 \right\}.$$

admet au moins une solution $u = (u_1, u_2)$ qui vérifient

$$\frac{\partial g^{**}}{\partial t_i}(x, Au) = p_i,$$

$$A_i p_i = \frac{\partial h_3}{\partial \eta_i}(x, u), \qquad i = 1, 2$$

où $p_i \in W^{2m,p}(\Omega)$. Comme dans le cas des problèmes $(\mathcal{P}_1)$ et $(\mathcal{P}_2)$, on montre à l'aide de (4) et (5) que l'ensemble

$$E = \{x \in \Omega / (x, Au(x)) \in K\}$$

est de mesure nulle, i.e., que l'on a

$$g_3^{**}(x, Au(x)) = g_3(x, Au(x)) \quad \text{p.p. } x \in \Omega,$$

montrant ainsi que u est solution de $(\mathcal{P}_3)$. □

Remerciement. Je remercie le référé pour ses remarques qui m'ont permis d'améliorer la présentation de ce travail.

REFERENCES

[1] R. C. ABEYARATNE, *Discontinuous deformation gradients in finite twisting of an incompressible elastic tube*, J. Elasticity, 11 (1981), pp. 43-80.

[2] S. S. ANTMAN, *Regular and singular problems for large elastic deformations of tubes, wedges, and cylinders*, Arch. Rational Mech. Anal., 83 (1983), pp. 1-52.

[3] G. AUBERT ET R. TAHRAOUI, *Théorèmes d'existence en calcul des variations*, J. Differential Equations, 33 (1979), pp. 1-15.

[4] ———, *Théorèmes d'existence en optimisation non convexe*, Appl. Anal., 18 (1984), pp. 75-100.

[5] ———, *Sur une classe de problèmes différentiels non linéaires par une méthode variationnelle*, Bollettino de l'U.M.I. (7) 3-B (1982), à paraître.

[6] J. M. BALL, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1977), pp. 337-403.

[7] P. G. CIARLET, *Mathematical Elasticity—Vol. I Three-Dimensional Elasticity*, Laboratoire d'Analyse Numérique, Université Parix VI, Paris, 1986.

[8] B. DACOROGNA, *Quasiconvexity and relaxation of nonconvex problems in the calculus of variations*, J. Funct. Anal., 46 (1982), pp. 102-118.

[9] I. EKELAND ET R. TEMAM, *Analyse convexe et problèmes variationnels*, Dunod, Gauthier-Villars, Paris, 1974.

[10] J. L. ERIKSEN, *Equilibrium of bars*, J. Elasticity, 5 (1975), pp. 191-201.

[11] R. L. FOSDICK ET R. D. JAMES, *The elastica and the problem of pure bending for nonconvex stored energy function*, J. Elasticity, 11 (1981), pp. 165-186.

[12] R. L. FOSDICK ET G. MACSITHIGH, *Helical shear of an elastic, circular tube with non-convex stored energy*, Arch. Rational Mech. Anal., 84 (1983), pp. 31-53.

[13] M. E. GURTIN, *Topics in Finite Elasticity*, CBMS-NSF Regional Conference Series in Applied Mathematics 35, Society for Industrial and Applied Mathematics, Philadelphia, 1981.

[14] J. K. KNOWLES ET E. STERNBERG, *Anti-plane shear fields with discontinuous deformation gradients near the tip of a crack in finite elastostatics*, J. Elasticity, 11 (1981), pp. 129-164.

[15] J. E. MARSDEN ET T. J. R. HUGHES, *Mathematical Foundations of Elasticity*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[16] R. S. RIVLIN, *Large elastic deformations of isotropic materials* VI. *Further results in the theory of torsion, shear and flexure*, Philos. Trans. Roy. Soc. London Ser. A, 242 (1949), pp. 173-195.

[17] R. TAHRAOUI, *Théorèmes d'existence en calcul des variations et applications à l'élasticité non linéaire*, Proc. Roy. Soc. Edinburgh, Sect. A., 109 (1988), pp. 51-78.

[18] ———, *Contribution à l'étude de quelques questions d'analyse non linéaire issues de la mécanique*, Thèse de doctorat, Université Paris VI, 1986.

[19] L. TARTAR, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics, Heriot-Watt Symposium Vol. IV, Heriot-Watt University, Edinburgh, U.K., 1978.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 53-75, January 1990
© 1990 Society for Industrial and Applied Mathematics
004

SEMIDISCRETIZATION METHOD FOR THREE-DIMENSIONAL MOTION OF A BINGHAM FLUID*

JONG UHN KIM†

Abstract. By the method of semidiscretization in a time variable, the existence of a strong solution to an initial boundary value problem associated with the three-dimensional motion of a Bingham fluid is established. The main tool used is a discretized version of the variation of constants formula combined with the L^p-theory of the Stokes operator.

Key words. Bingham fluid, semidiscretization method, variation of constants formula, L^p-theory of the Stokes operator

AMS(MOS) subject classifications. 35B10, 35B65, 35K55, 76D99

0. Introduction. In this paper, we present a new result on the existence of solutions of an initial-boundary value problem associated with the motion of a Bingham fluid in a three-dimensional domain. A Bingham fluid is a rigid viscoplastic fluid that is a particular kind of non-Newtonian fluid. This material behaves in a rigid manner when a certain function of stresses does not reach the yield limit; it moves like a Newtonian fluid beyond this limit.

Common examples of Bingham fluids are slurries, drilling muds, oil paints, and toothpaste. Engineering applications (particularly in the chemical and process industries), such as experimental techniques, rolling and extrusion processes, and heat transfer, are discussed in [16], where a list of engineering references will also be found. For the continuum mechanics foundations, see [13].

Since the relation between strains and stresses becomes very different depending on the state of stresses, the motion of a Bingham fluid cannot be described by a single equation. This difficulty was overcome by Duvaut and Lions, who derived a variational inequality that can take care of the unknown interface between rigid medium and fluid zone [3], [4]. They formulated an initial-boundary value problem for a Bingham fluid as follows:

$$(0.1)\qquad (\partial u/\partial t,\, w-u)+a(u,\, w-u)+b(u,\, u,\, w)+J(w)-J(u)\geqq (f,\, w-u)\ \text{in}\ (0,\, T),$$

for each test function w such that $\nabla\cdot w=0$ in Ω and $w=0$ on $\partial\Omega$,

$$(0.2)\qquad \nabla\cdot u=0 \quad \text{in}\ \Omega\times(0,\, T),$$

$$(0.3)\qquad u=0 \quad \text{on}\ \partial\Omega\times[0,\, T],$$

$$(0.4)\qquad u(x,\, 0)=u_0(x) \quad \text{in}\ \Omega.$$

Here, Ω is a bounded domain in $\mathbb{R}^3$ with C^∞ boundary $\partial\Omega$, $u(x,\, t)$ denotes the velocity of the fluid, and $f(x,\, t)$ stands for external force. The density, the yield limit, and the viscosity are assumed to be positive constants. In particular, the density is taken to be

* Received by the editors October 5, 1987; accepted for publication (in revised form) February 2, 1989. This research was supported by Air Force Office of Scientific Research grant AFOSR-86-0085 and by National Science Foundation grant DMS-8521848.

† Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

1. We employ the notation:

$$a(u, w) = \sum_{i,j=1}^{3} 2\mu \int_{\Omega} D_{ij}(u) D_{ij}(w)\, dx, \qquad \mu = \text{viscosity},$$

$$D_{ij}(u) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right),$$

$$J(u) = 2g \int_{\Omega} D_{\Pi}(u)^{1/2}\, dx, \qquad g = \text{yield limit},$$

$$D_{\Pi} = \frac{1}{2} \sum_{i,j=1}^{3} D_{ij}(u)^2,$$

$$b(u, v, w) = \sum_{i,j=1}^{3} \int_{\Omega} u_j \frac{\partial v_i}{\partial x_j} w_i\, dx.$$

$(,)$ is a scalar product that will be defined in the next section.

The conservation of momentum is expressed by (0.1) and the condition of incompressibility is given by (0.2).

It is easy to see that (0.1) reduces to the Navier–Stokes equations if the yield limit g vanishes. The nondifferentiable functional $J(\cdot)$ causes a serious mathematical difficulty in addition to the difficulties inherited from the Navier–Stokes equations. This makes the initial boundary value problem very challenging. In the case of laminar flow in a cylindrical pipe, the problem simplifies to a special case, for which the mathematical analysis is complete. This special case also has been used as a typical model in the finite-element analysis of parabolic variational inequalities (see [8] and references therein). We focus on the existence of solutions to the more general problem in a three-dimensional domain.

We will survey known results that have motivated the present investigation.

Duvaut and Lions [3] have proved for the first time the existence of weak solutions of (0.1)–(0.4). The weak solutions according to the definition in [3] and [4] belong to the same function class as the Leray–Hopf weak solutions of the Navier–Stokes equations. For a three-dimensional domain, these weak solutions satisfy a weak version of (0.1) and the uniqueness of weak solutions is an open question.

It is known that for smooth data, there is a local strong solution to the initial-boundary value problem for the Navier–Stokes equations. For the Cauchy problem associated with (0.1) in R^2 or R^3, it is known that strong solutions (local in time in R^3) exist if the data are regular. The result is the same as that for the Navier–Stokes equations (see [9], [14]). For the initial-boundary value problem shown above, a different kind of strong solution was obtained in [3] and [4] in the case of a two-dimensional domain. An analogous result was established in [10] for a three-dimensional domain by assuming that the initial data are stationary states with external force in $L^2(\Omega)^3$. Under the same assumption on the data, Naumann and Wulst [12] obtained a similar result for a variant of a Bingham fluid through a different method. This assumption on the initial data is not satisfied in general even by C^∞ divergence-free vector fields with compact support. Hence, this is not an ordinary regularity assumption and is very restrictive.

The purpose of the present paper is to eliminate this assumption on the initial data. We show that if the initial data are divergence-free vector fields that belong to $L^r(\Omega)^3$, $r > 3$, and whose normal component vanishes on the boundary, there indeed exists at least a local solution regular enough to be unique (see Theorem 2.2). The result is comparable to that of Giga and Miyakawa [7] for the Navier–Stokes equations.

The basic approach for the problem (0.1)-(0.4), initiated by Duvaut and Lions, consists of three steps. The first step is to reduce inequality (0.1) to an equation by substituting a differentiable function with regularizing parameter for $J(\cdot)$. The second step is to obtain a solution for the regularized problem together with uniform estimates independent of the regularizing parameter. The third step is to prove that the limit of a sequence of approximate solutions exists and is a solution of the original problem.

In the second step, we are tempted to use one of two well-known techniques for the Navier-Stokes equations. One such technique is the Galerkin approximation method. For our problem, it seems that we cannot obtain enough estimates independent of the regularizing parameter through this method unless special assumptions are made on the initial data as in [10]. The other method, used in [7], is to set up the variation of the constants formula involving the analytic semigroup generated by the Stokes operator in $L^r(\Omega)^3$ and then to employ the related iteration scheme to find a solution. When we attempt to apply this method to our problem, serious technical difficulties arise. These difficulties can be avoided, however, if we use a discretized version of the variation of the constants formula. The method we will use is basically the semidiscretization method used in [15] to give an alternate proof of the existence of the Leray-Hopf weak solution to the Navier-Stokes equations. While our scheme of discretization is a slight modification of that used in [15], we obtain substantially better estimates by interpreting the scheme as a discretized version of the variation of the constants formula, and we retain all the basic L^2 estimates of [15]. The estimates obtained through this procedure are sufficient for the pointwise convergence of approximating solutions.

Finally we also prove the existence of global solutions and time-periodic solutions by assuming that the initial data and external force are sufficiently small; see Theorems 3.1 and 3.2.

1. Notation and preliminaries. Throughout this paper, Ω denotes a bounded domain in R^3 with C^∞ boundary. We employ the notation defined in the Introduction, as well as the following:

$$\partial_t = \frac{\partial}{\partial t}, \quad \partial_i = \frac{\partial}{\partial x_i} \quad \text{for } 1 = 1, 2, 3, \quad \Delta = \sum_{i=1}^{3} \partial_i^2, \quad \nabla = (\partial_1, \partial_2, \partial_3),$$

$$\nabla \cdot f = \sum_{i=1}^{3} \partial_i f_i \quad \text{for } f = (f_1, f_2, f_3).$$

When E is a Banach space, $L^r(0, T; E)$ is the set of all E-valued strongly measurable L^r functions on $[0, T]$ with the obvious norm. $C(I; E)$ is the set of all E-valued continuous functions on the interval I.

We introduce the following function spaces:

$S = \{\phi \in C_0^\infty(\Omega)^3 : \nabla \cdot \phi = 0 \text{ in } \Omega\}$,

$W^{m,r}(\Omega) = \{v \in L^r(\Omega) : \partial_1^{\alpha_1}\partial_2^{\alpha_2}\partial_3^{\alpha_3} v \in L^r(\Omega), 1 \leqq \alpha_1 + \alpha_2 + \alpha_3 \leqq m\}$,

$W_0^{m,r}(\Omega) =$ the completion of $C_0^\infty(\Omega)$ in $W^{m,r}(\Omega)$,

$W^{-m,r'}(\Omega) =$ the dual of $W_0^{m,r}(\Omega)$, where $1/r' + 1/r = 1$, $1 \leqq r < \infty$,

$X_r =$ the completion of S in $L^r(\Omega)^3$, $1 < r < \infty$,

$V = W_0^{1,2}(\Omega)^3 \cap X_2$,

$V' =$ the dual of V.

$(\ ,\ )$ stands for the duality pairing between V and V'. In particular, if $v \in X_2$ and $w \in V$, then (v, w) coincides with the scalar product of v and w in X_2. We can characterize X_r by

$$X_r = \{v \in L^r(\Omega)^3 : \nabla \cdot v = 0 \text{ in } \Omega \text{ and the normal component of } v \text{ vanishes on } \partial\Omega\}.$$

We let P_r denote the projection from $L^r(\Omega)^3$ onto X_r and write the Stokes operator as

$$A_r = -P_r\Delta \quad \text{for } 1<r<\infty,$$

with the domain

$$\mathcal{D}(A_r) = W^{2,r}(\Omega)^3 \cap W_0^{1,r}(\Omega)^3 \cap X_r.$$

We list some basic properties of A_r. Giga [5] proved that for any $\varepsilon>0$ and $1<r<\infty$, there is a positive constant $C_{\varepsilon,r}$ such that

$$\|(\lambda I + A_r)^{-1}\| \leqq C_{\varepsilon,r}/|\lambda| \tag{1.1}$$

for all $\lambda \in \mathbb{C}$ such that $\lambda \neq 0$ and $|\arg \lambda| \leqq \pi - \varepsilon$, where $\|\cdot\|$ is the norm of a bounded linear operator from X_r into X_r. We also note that zero belongs to the resolvent set of A_r.

Using (1.1), we can define A_r^θ, $0<\theta<1$ by the Dunford integral and the domain of A_r^θ is defined by complex interpolation:

$$\mathcal{D}(A_r^\theta) = [X_r, \mathcal{D}(A_r)]_\theta, \qquad 0 \leqq \theta \leqq 1$$

(see [6]). Since $\mathcal{D}(A_r)$ is compactly embedded into X_r, we have Lemma 1.1.

LEMMA 1.1. *For $0 \leqq \theta_1 < \theta_2 \leqq 1$, the embedding $\mathcal{D}(A_r^{\theta_2}) \subset \mathcal{D}(A_r^{\theta_1})$ is compact.*

Next we obtain some estimates to be used later.

LEMMA 1.2. *Let $1<r<\infty$ and $0 \leqq \alpha \leqq 1$. Then, for any integer $k \geqq 1$ and any $0<\varepsilon \leqq 1$,*

$$\|A_r^\alpha (I+\varepsilon A_r)^{-k}\| \leqq C_{r,\alpha} \exp(-\delta\varepsilon k)/(\varepsilon k)^\alpha, \tag{1.2}$$

where $C_{r,\alpha}$ and δ are positive constants independent of ε and k.

Proof. By virtue of (1.1) and the fact that the resolvent set of A_r contains zero, we can write, for any positive number λ and any integer $k \geqq 2$,

$$-A_r(\lambda I + A_r)^{-k} = \frac{1}{2\pi i}\int_\Gamma z(\lambda - z)^{-k}(zI + A_r)^{-1}\, dz, \tag{1.3}$$

where $\Gamma = \{-\xi + \rho e^{i\theta}: 0 \leqq \rho < \infty\} \cup \{-\xi + \rho e^{-i\theta}: 0 \leqq \rho < \infty\}$ and we choose $\xi > 0$, $\pi/2 < \theta < \pi$ and the orientation of Γ so that Γ is contained in the resolvent set of A_r and $\operatorname{Im}\Gamma$ is increasing along Γ.

It follows from (1.1) and (1.3) that

$$\begin{aligned} \|A_r(\lambda I + A_r)^{-k}\| &\leqq M\int_0^\infty (\lambda + \xi + \rho|\cos\theta|)^{-k}\, d\rho, \\ &\leqq M(\lambda+\xi)^{-k+1}/(k-1), \end{aligned} \tag{1.4}$$

where M denotes positive constants independent of λ and k. Hence, by setting $\varepsilon = 1/\lambda$ and assuming $0<\varepsilon \leqq 1$, we derive

$$\begin{aligned} \|A_r(I+\varepsilon A_r)^{-k}\| &\leqq M(1+\varepsilon\xi)^{-k+1}/\varepsilon(k-1) \\ &\leqq M \exp(-\delta_1\varepsilon k)/(\varepsilon k), \end{aligned} \tag{1.5}$$

where M and δ_1 denote positive constants independent of ε and k. The second inequality in (1.5) is also true for $k=1$, which is easily seen by virtue of (1.1). Since A_r is the infinitesimal generator of a bounded analytic semigroup and the resolvent set of A_r contains zero, it is known that for any integer $k \geqq 1$,

$$\|(\lambda I + A_r)^{-k}\| \leqq M(\lambda + \tilde{\xi})^{-k}, \tag{1.6}$$

where M and $\tilde{\xi}$ are positive constants independent of λ and k. Hence, by setting $\varepsilon = 1/\lambda$, we have

$$\|(I+\varepsilon A_r)^{-k}\| \leqq M(1+\varepsilon\tilde{\xi})^{-k}, \tag{1.7}$$

which, together with the assumption $0<\varepsilon\leqq 1$, yields

$$\|(I+\varepsilon A_r)^{-k}\| \leqq M\exp(-\delta_2\varepsilon k), \tag{1.8}$$

where M and δ_2 are positive constants independent of ε and k. For $0<\alpha<1$, we obtain (1.2) from (1.5) and (1.8) by means of the interpolation inequality.

LEMMA 1.3. *Let $1<r<\infty$ and $0<\alpha\leqq 1$. Then, for any integer $k\geqq 1$ and any $\varepsilon>0$, we have*

$$\|((I+\varepsilon A_r)^{-k}-I)A_r^{-\alpha}\| \leqq C_\alpha(\varepsilon k)^\alpha, \tag{1.9}$$

where C_α is a positive constant independent of ε and k.

Proof. We first note that

$$(I+\varepsilon A_r)^{-1}-I = -\varepsilon A_r(I+\varepsilon A_r)^{-1}, \tag{1.10}$$

which, combined with (1.7), yields

$$\|((I+\varepsilon A_r)^{-1}-I)A_r^{-1}\| \leqq \varepsilon M, \tag{1.11}$$

where M is a positive constant independent of ε. Following [2], we write for $k\geqq 2$

$$((I+\varepsilon A_r)^{-k}-I)A_r^{-1} = ((I+\varepsilon A_r)^{-1}-I)A_r^{-1} + \sum_{j=1}^{k-1}(I+\varepsilon A_r)^{-j}((I+\varepsilon A_r)^{-1}-I)A_r^{-1}. \tag{1.12}$$

Again by (1.7), we derive from (1.12)

$$\|((I+\varepsilon A_r)^{-k}-I)A_r^{-1}\| \leqq M\varepsilon k, \tag{1.13}$$

where M is a positive constant independent of ε and k. Inequality (1.9) is now a simple consequence of (1.7), (1.11), and (1.13) through the interpolation. □

LEMMA 1.4. *Let $v\in X_r$, $1<r<\infty$. Then, for any given $\xi>0$, there is $\delta(\xi, v)>0$ such that for all $\varepsilon>0$ and $k\geqq 1$ satisfying $\varepsilon k\leqq\delta(\xi, v)$,*

$$\|((I+\varepsilon A_r)^{-k}-I)v\|_{X_r} \leqq \xi. \tag{1.14}$$

Proof. Suppose that the assertion above is false. Then there are $\xi>0$, $\{\varepsilon_n\}_{n=1}^\infty$ and $\{k_n\}_{n=1}^\infty$ such that $\varepsilon_n>0$, $k_n\geqq 1$, $\varepsilon_n k_n\to 0$ as $n\to\infty$, and

$$\|((I+\varepsilon_n A_r)^{-k_n}-I)v\|_{X_r} > \xi \quad \text{for all } n. \tag{1.15}$$

It follows from (1.13) that for all $w\in\mathcal{D}(A_r)$,

$$\|((I+\varepsilon_n A_r)^{-k_n}-I)w\|_{X_r} \leqq M\varepsilon_n k_n\|A_r w\|_{X_r} \tag{1.16}$$

where M is a positive constant independent of ε_n, k_n, and w.

In the meantime, (1.7) implies that

$$\|(I+\varepsilon_n A_r)^{-k_n}-I\| \leqq \tilde{M} \quad \text{for all } \varepsilon_n \text{ and } k_n. \tag{1.17}$$

We now choose $w\in\mathcal{D}(A_r)$ such that

$$\|v-w\|_{X_r} \leqq \xi/2\tilde{M}. \tag{1.18}$$

Then we find that

$$\begin{aligned}\|((I+\varepsilon_n A_r)^{-k_n}-I)v\|_{X_r} &\leqq \|((I+\varepsilon_n A_r)^{-k_n}-I)(w-v)\|_{X_r} + M\varepsilon_n k_n\|A_r w\|_{X_r}\\ &\leqq \tfrac{1}{2}\xi + M\varepsilon_n k_n\|A_r w\|_{X_r},\end{aligned} \tag{1.19}$$

which contradicts (1.15) as $\varepsilon_n k_n\to 0$. This concludes the proof. □

We list some properties of the operator $A_r^{\beta} P_r \partial_j$, $j=1,2,3$, that are proved in [7]. Let $1<r<\infty$. Then there is a positive constant C_r such that

$$\|A_r^{-1/2} P_r \partial_j v\|_{X_r} \leqq C_r \|v\|_{L^r(\Omega)^3}, \qquad j=1,2,3, \tag{1.20}$$

for all $v \in W^{1,r}(\Omega)^3$. Hence, $A_r^{-1/2} P_r \partial_j$ can be extended to a bounded linear operator from $L^r(\Omega)^3$ into X_r, $1<r<\infty$. If $3<r<\infty$, $6r/(3+r)<p \leqq r$ and $\delta = \frac{1}{2}+\frac{3}{2}(2/p-1/r)$, then there is a constant $C_{r,p}$ such that

$$\|A_r^{-\delta} P_r \partial_j v\|_{X_r} \leqq C_{r,p} \|v\|_{L^{p/2}(\Omega)^3} \quad \text{for all } v \in W^{1,r}(\Omega)^3. \tag{1.21}$$

Hence, $A_r^{-\delta} P_r \partial_j$ can be extended to a bounded linear operator from $L^{p/2}(\Omega)^3$ into $X_r(\Omega)$, for $3<r<\infty$, $6r/(3+r)<p \leqq r$. If $3<r<\infty$, $0<\nu<3/2r$ and $\delta = 3/2r+\frac{1}{2}-2\nu$, then there is a constant C_r such that for all v, $w \in \mathscr{D}(A_r)$,

$$\|A_r^{-\delta} P_r \sum_{j=1}^{3} \partial_j (v_j w)\|_{X_r} \leqq C_r \|A_r^{\nu} v\|_{X_r} \|A_r^{\nu} w\|_{X_r}, \tag{1.22}$$

where $v=(v_1, v_2, v_3)$.

Hence, the mapping $(v, w) \to A_r^{-\delta} P_r \sum_{j=1}^{3} \partial_j (v_j w)$ can be extended to a bounded bilinear mapping from $\mathscr{D}(A_r^{\nu})^2$ into X_r.

Next we consider the boundary value problem:

$$u - \varepsilon \Delta u + \nabla p = h \quad \text{in } \Omega, \qquad \varepsilon > 0, \tag{1.23}$$

$$\nabla \cdot u = 0 \quad \text{in } \Omega, \tag{1.24}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{1.25}$$

It is understood that (1.23) and (1.24) hold in the sense of distribution in Ω. If $h = \partial_j v$ and $v \in L^r(\Omega)^3$, $1<r<\infty$, then the unique solution of the problem above in $W_0^{1,r}(\Omega)^3$ can be expressed as

$$u = A_r^{1/2} (I + \varepsilon A_r)^{-1} A_r^{-1/2} P_r \partial_j v. \tag{1.26}$$

Suppose that $3<r<\infty$, $6r/(3+r)<p \leqq r$, $\delta = \frac{1}{2}+\frac{3}{2}(2/p-1/r)$ and that $h = \sum_{j=1}^{3} \partial_j (v_j w)$, where $v=(v_1, v_2, v_3)$ and $v, w \in X_p$. Then, the unique solution in $W_0^{1,p/2}(\Omega)^3$ can be expressed as

$$u = A_r^{\delta} (I + \varepsilon A_r)^{-1} A_r^{-\delta} P_r \sum_{j=1}^{3} \partial_j (v_j w). \tag{1.27}$$

If $3<r<\infty$, $0 \leqq \nu < 3/2r$, $\delta = 3/2r+\frac{1}{2}-2\nu$, and $h = \sum_{j=1}^{3} \partial_j (v_j w)$ with $v, w \in \mathscr{D}(A_r^{\nu})$, then the unique solution in $W_0^{1,r/2}(\Omega)^3$ can be expressed as

$$u = A_r^{\delta} (I + \varepsilon A_r)^{-1} A_r^{-\delta} P_r \sum_{j=1}^{3} \partial_j (v_j w). \tag{1.28}$$

These expressions will be used in the subsequent section. We prove (1.26). Let $h = \partial_j v$ and $v \in L^r(\Omega)^3$, $1<r<\infty$. We choose a sequence $\{v_n\}_{n=1}^{\infty}$ in $W_0^{1,r}(\Omega)^3$ such that v_n converges to v strongly in $L^r(\Omega)^3$. For each n, we write

$$u_n = (I + \varepsilon A_r)^{-1} P_r \partial_j v_n. \tag{1.29}$$

Then $u_n \in \mathscr{D}(A_r)$ and u_n satisfies

$$u_n - \varepsilon \Delta u_n + \nabla p_n = \partial_j v_n \quad \text{in } \Omega \tag{1.30}$$

for some function $p_n \in W^{1,r}(\Omega)$.

We now rewrite (1.29) as

$$u_n = A_r^{1/2} (I + \varepsilon A_r)^{-1} A_r^{-1/2} P_r \partial_j v_n. \tag{1.31}$$

Since $A_r^{-1/2}P_r\partial_j$ is a bounded linear operator from $L^r(\Omega)^3$ into X_r and $A_r^{1/2}(I+\varepsilon A_r)^{-1}$ is a bounded linear operator from X_r into $\mathcal{D}(A_r^{1/2})$, $\{u_n\}_{n=1}^{\infty}$ converges strongly in $\mathcal{D}(A_r^{1/2})$ to

$$u = A_r^{1/2}(I+\varepsilon A_r)^{-1}A_r^{-1/2}P_r\partial_j v. \tag{1.32}$$

From (1.30), we find that u above satisfies

$$u - \varepsilon\Delta u + \nabla p = \partial_j v \tag{1.33}$$

in the sense of distribution in Ω for some scalar function p. Hence, u is a solution of (1.23)-(1.25). The uniqueness of u in $W_0^{1,r}(\Omega)^3$ follows by the duality argument and the existence of the solution in $W_0^{1,r'}(\Omega)^3$, $1/r'+1/r=1$, when $h \in W^{-1,r'}(\Omega)^3$. Now the proof of (1.26) is complete. □

We proceed to prove (1.27). Choose $\{v_n\}_{n=1}^{\infty}$ and $\{w_n\}_{n=1}^{\infty}$ in S such that v_n and w_n converge to v and w, respectively, strongly in X_p. It is apparent that $v_{nj}w_n \in W_0^{1,r}(\Omega)^3$, for $j=1,2,3$, where $v_n=(v_{n1}, v_{n2}, v_{n3})$, and thus, $P_r\partial_j(v_{nj}w_n) = P_{p/2}\partial_j(v_{nj}w_n)$, for $j=1,2,3$.

As above, we write for each n

$$u_n = \sum_{j=1}^{3} A_{p/2}^{1/2}(I+\varepsilon A_{p/2})^{-1}A_{p/2}^{-1/2}P_{p/2}\partial_j(v_{nj}w_n), \tag{1.34}$$

and note that u_n converges to u strongly in $\mathcal{D}(A_{p/2}^{1/2})$, where u is the unique solution in $W_0^{1,p/2}(\Omega)^3$ of (1.23)-(1.25) with $h=\sum_{j=1}^{3}\partial_j(v_j w)$. Rewriting (1.34) as

$$\begin{aligned} u_n &= \sum_{j=1}^{3} (I+\varepsilon A_{p/2})^{-1}P_{p/2}\partial_j(v_{nj}w_n) \\ &= \sum_{j=1}^{3} (I+\varepsilon A_r)^{-1}P_r\partial_j(v_{nj}w_n) \\ &= \sum_{j=1}^{3} A_r^{\delta}(I+\varepsilon A_r)^{-1}A_r^{-\delta}P_r\partial_j(v_{nj}w_n), \end{aligned} \tag{1.35}$$

we derive (1.27) from (1.21) and the fact that u_n converges to u strongly in $\mathcal{D}(A_{p/2}^{1/2})$. The proof of (1.28) is very similar and we omit it.

We use a regularized version of $J(\cdot)$,

$$J_\eta(v) = 2g\int_\Omega (\eta + D_{\mathrm{II}}(v))^{1/2}\,dx, \tag{1.36}$$

where η is a positive number and g is the yield limit. The Gâteaux differential $J_\eta(\cdot)$ is given by

$$(J_\eta'(v), w) = g\int_\Omega \sum_{i,j=1}^{3} (\eta + D_{\mathrm{II}}(v))^{-1/2}D_{ij}(v)D_{ij}(w)\,dx \tag{1.37}$$

for each v, $w \in V$. Since $J_\eta(\cdot)$ is convex, $J_\eta'(\cdot)$ is monotone and

$$J_\eta(v) - J_\eta(w) \geqq (J_\eta'(w), v-w) \tag{1.38}$$

for all v, $w \in V$. We also use the inequality

$$\|(\eta + D_{\mathrm{II}}(v))^{-1/2}D_{ij}(v)\|_{L^\infty(\Omega)} \leqq \sqrt{2} \quad \text{for } i,j=1,2,3 \tag{1.39}$$

for every $\eta > 0$ and $v \in W^{1,1}(\Omega)^3$.

2. Local existence and uniqueness of solutions.

DEFINITION 2.1. A function $u(x, t)$ is called a solution (0.1)-(0.4) on an interval $[0, T)$ if

(i) $u \in L^2(0, T; V)$ and $\partial_t u \in L^2(0, T; V')$;

(ii) (0.1) is satisfied for every $w \in V$, for almost all $t \in (0, T)$;

(iii) $u(x, 0) = u_0(x)$.

Condition (iii) makes sense since (i) implies $u \in C([0, T]; X_2)$, possibly after a modification on a set of measure zero.

This definition of solution is stronger than that of weak solution in [3] and [4]. Our main result is Theorem 2.2.

THEOREM 2.2. *Suppose that* $3 < r < \infty$, $u_0(x) \in X_r$, *and* $f \in L^\infty(0, T; W^{-1,r}(\Omega)^3)$. *Then there is a unique solution* $u(x, t)$ *on an interval* $[0, T^*)$, *where* $0 < T^* \leqq T$. *Furthermore,* $u \in C([0, T^*); X_r)$ *and, for each* $0 < \delta \leqq \frac{1}{2}$, u *is* $\mathscr{D}(A_r^{1/2-\delta})$*-valued locally Hölder continuous on* $(0, T^*)$.

In fact, the assertion above is a consequence of Theorem 2.3.

THEOREM 2.3. *Let* $u_0(x) \in \mathscr{D}(A_r^\nu)$, $f \in L^\infty(0, T; W^{-1,r}(\Omega)^3)$, $0 \leqq \nu \leqq 3/4r$, *and* $3 < r < \infty$. *Then, there is a unique solution* $u(x, t)$ *on an interval* $[0, T_1)$, *where* $0 < T_1 \leqq T$, *and* T_1 *depends only on* ν, r, $\|u_0\|_{\mathscr{D}(A_r^\nu)}$ *and* $\|f\|_{L^\infty(0,T;W^{-1,r}(\Omega)^3)}$. *In particular,* T_1 *is nonincreasing in* $\|u_0\|_{\mathscr{D}(D_r^\nu)}$. *Furthermore,*

$$u \in C([0, T_1]; \mathscr{D}(A_r^\nu)), \tag{2.1}$$

and, for each $0 < \rho < T_1$ *and each* $0 \leqq \alpha < \alpha + \beta < \frac{1}{2} - 3/2r + \nu$,

$$u \in C((0, T_1]; \mathscr{D}(A_r^{\nu+\alpha})), \tag{2.2}$$

$$\|u(s_2) - u(s_1)\|_{\mathscr{D}(A_r^{\nu+\alpha})} \leqq M(s_2 - s_1)^\beta \quad \text{for all } 0 < \rho \leqq s_1 < s_2 < T_1, \tag{2.3}$$

where M *is a positive constant depending on* $r, \rho, \alpha, \beta, \nu$, $\|u_0\|_{\mathscr{D}(A_r^\nu)}$, *and* $\|f\|_{L^\infty(0,T;W^{-1,r}(\Omega)^3)}$.

We will show that Theorem 2.3 implies Theorem 2.2. Let us fix $3 < r < \infty$, $u_0(x) \in X_r$, and $f \in L^\infty(0, T; W^{-1,r}(\Omega)^3)$. Then, by Theorem 2.3 with $\nu = 0$, there is a unique solution u on an interval $[0, T_1)$, and u satisfies (2.1)-(2.3) with $\nu = 0$. Next we show that u is $\mathscr{D}(A_r^{1/2-\delta})$-valued, locally Hölder continuous on $(0, T_1]$ for each $0 < \delta \leqq \frac{1}{2}$. We use the uniqueness of solution and the fact that the solution becomes more regular than the initial datum.

Without loss of generality, let us assume $0 < T_1 < T$. Choose any $0 < \xi < T_1/2$. Then, $u \in C([\xi, T_1]: \mathscr{D}(A_r^\lambda))$ with $\lambda = \min(3/4r, \frac{1}{2}(\frac{1}{2} - 3/2r))$ and

$$\|u\|_{\mathscr{D}(A_r^\lambda)} \leqq M \quad \text{for all } t \in [\xi, T_1] \tag{2.4}$$

for some positive constant M.

For any $s \in [\xi, T_1]$, we can apply Theorem 2.3 with s as initial time, $u(s)$ as an initial datum, and $\nu = \lambda$. We then obtain a solution

$$v_s \in C([s, s+h]; \mathscr{D}(A_r^\lambda)) \cap C([s+\tau, s+h]; \mathscr{D}(A_r^{2\lambda})), \tag{2.5}$$

where $h > 0$ can be chosen independently of s by (2.4) and $\tau = \min(h/2, \xi/2)$. By the uniqueness of solutions, $v_s = u$ on $[s, s+h]$. Hence we can derive that

$$u \in C\left(\left[\xi + \frac{\xi}{2}, T_1\right]; \mathscr{D}(A_r^{2\lambda})\right). \tag{2.6}$$

If $\frac{9}{2} \leqq r$, (2.6) implies

$$u \in C((0, T_1]; \mathscr{D}(A_r^{3/4r})). \tag{2.7}$$

If $3<r<\frac{9}{2}$, we can repeat the same argument until we arrive at (2.7). Next choose any $0<\delta<\frac{1}{2}$ and $0<\rho<T_1/2$. Using $t=\rho/2$ as initial time and $u(\rho/2)$ as an initial datum, we can apply Theorem 2.3 with $\nu=3/4r$ and $\alpha=\max(0, \frac{1}{2}-3/4r-\delta)$ to conclude that u is $\mathscr{D}(A_r^{1/2-\delta})$-valued locally Hölder continuous on $(0, T_1]$. Finally we choose

$$T^*=\sup\{\tilde{T}: T_1\leqq\tilde{T}\leqq T \text{ and there is a solution that is } X_r\text{-valued continuous on } [0, \tilde{T}]\}.$$

By repeating the above argument, we find that u is $\mathscr{D}(A_r^{1/2-\delta})$-valued locally Hölder continuous on $(0, T^*)$. This ends the proof of Theorem 2.2. □

We outline the strategy of proof of Theorem 2.3.

Step 1. We replace $J(\cdot)$ by $J_\eta(\cdot)$. Using the differential $J_\eta'(\cdot)$ and discretizing time variable, we construct a sequence of approximate solutions that are piecewise linear in time. The standard L^2-estimates are obtained as in the case of the Navier-Stokes equations (see [15]). With the aid of the properties of the Stokes operator in L^r, we also obtain L^r estimates independent of η and the meshsize in time variable.

Step 2. It is shown that this sequence of approximate solutions converges to a solution of (0.1)-(0.4) with $J(\cdot)$ replaced by $J_\eta(\cdot)$ as the meshsize tends to zero.

Step 3. We pass η to zero to obtain a solution according to Definition 2.1 and prove the uniqueness of solution.

Throughout this section, we fix $3<r<\infty$ and use the notation

$$A=\mu A_r \quad \text{and} \quad P=P_r.$$

We also suppose that $0\leqq\nu\leqq 3/4r$, $u_0(x)\in\mathscr{D}(A^\nu)$, and $f\in L^\infty(0, T; W^{-1,r}(\Omega)^3)$ are given.

2.1. Construction and estimates of approximate solutions. We choose any positive integer N and set

$$\varepsilon=\frac{T}{N}. \tag{2.8}$$

Define an approximate solution W_N by

$$W_N=\frac{t-k\varepsilon}{\varepsilon}(u_{k+1}-u_k)+u_k \quad \text{for } k\varepsilon\leqq t\leqq(k+1)\varepsilon, \quad k=0, 1, \cdots, N-1, \tag{2.9}$$

where u_0 is the given initial datum and u_{k+1} is determined from the equation

$$(u_{k+1}, \phi)+\varepsilon a(u_{k+1}, \phi)+\varepsilon b(u_k, u_{k+1}, \phi)+\varepsilon(J_\eta'(u_{k+1}), \phi)=(u_k, \phi)+\varepsilon(f_{k+1}, \phi) \tag{2.10}$$

for all $\phi\in S$, $k=0, 1, \cdots, N-1$, where $f_{k+1}=1/\varepsilon\int_{k\varepsilon}^{(k+1)\varepsilon} f\,dt$, for $k=0, 1, \cdots, N-1$, and $(,)$ denotes the duality pairing between V and V'.

LEMMA 2.4. *If $u_k\in X_r$ and $f_{k+1}\in W^{-1,r}(\Omega)^3$, then there is a unique solution u_{k+1} of* (2.10) *in $V\cap W_0^{1,r}(\Omega)^3$.*

Proof. Fix any $u_k\in X_r$ and define a mapping Λ from V into its dual V' such that for every $v, w\in V$,

$$(\Lambda v, w)=(v, w)+\varepsilon a(v, w)+\varepsilon b(u_k, v, w)+\varepsilon(J_\eta'(v), w). \tag{2.11}$$

Since $J_\eta'(\cdot)$ is monotone, it is easily seen that Λ is monotone. Furthermore, Λ is bounded, hemicontinuous and

$$\frac{(\Lambda v, v)}{\|v\|_V}\to\infty \quad \text{as } \|v\|_V\to\infty. \tag{2.12}$$

By virtue of Theorem 2.1 of [11, p. 171], Λ is surjective. Since f_{k+1} can be regarded as

an element of V', there is a function $u_{k+1} \in V$ such that

$$(\Lambda u_{k+1}, w) = (u_k + \varepsilon f_{k+1}, w) \quad \text{for all } w \in V, \tag{2.13}$$

which implies (2.10). Uniqueness follows from the strict monotonicity of Λ. Equation (2.13) implies that

$$\begin{aligned} u_{k+1} - \varepsilon\mu \Delta u_{k+1} + \varepsilon \sum_{j=1}^{3} u_{kj}\partial_j u_{k+1} - \varepsilon g \sum_{j=1}^{3} \partial_j \{(\eta + D_{\Pi}(u_{k+1}))^{-1/2} D_{ij}(u_{k+1})\} \\ + \nabla p = u_k + \varepsilon f_{k+1} \end{aligned} \tag{2.14}$$

holds in the sense of distribution in Ω for some scalar function p, where $u_k = (u_{k1}, u_{k2}, u_{k3})$ and the fourth term is a vector function represented by its ith component.

It remains to prove $u_{k+1} \in W_0^{1,r}(\Omega)^3$. Since $u_{k+1} \in V \subset W_0^{1,2}(\Omega)^3$, it follows that $u_{k+1} \in L^6(\Omega)^3$, and thus $\sum_{j=1}^{3} u_{kj}\partial_j u_{k+1} \in W^{-1,6r/(6+r)}(\Omega)^3$. This, combined with (1.39), yields $u_{k+1} \in W_0^{1,6r/(6+r)}(\Omega)^3$ (see [1]). If $r > 6$, then $W_0^{1,6r/(6+r)}(\Omega) \subset L^{\infty}(\Omega)$, which implies $\sum_{j=1}^{3} u_{kj}\partial_j u_{k+1} \in W^{-1,r}(\Omega)^3$, and hence $u_{k+1} \in W_0^{1,r}(\Omega)^3$. If $r = 6$, then $\sum_{j=1}^{3} u_{kj}\partial_j u_{k+1} \in W^{-1,3}(\Omega)^3$. Consequently, $u_{k+1} \in W_0^{1,3}(\Omega)^3$, which yields $\sum_{j=1}^{3} u_{kj}\partial_j u_{k+1} \in W^{-1,q}(\Omega)^3$, for any $1 \leqq q < 6$. It now follows that $u_{k+1} \in W_0^{1,q}(\Omega)^3$, for any $1 \leqq q < 6$. Thus, $u_{k+1} \in L^{\infty}(\Omega)^3$, which gives $u_{k+1} \in W_0^{1,6}(\Omega)^3$ by the same argument. If $3 < r < 6$, then $W_0^{1,6r/(6+r)}(\Omega) \subset L^{6r/(6-r)}(\Omega)$, and hence $\sum_{j=1}^{3} u_{kj}\partial_j u_{k+1} \in W^{-1,6r/(12-r)}(\Omega)^3$, which yields $u_{k+1} \in W_0^{1,6r/(12-r)}(\Omega)^3$. By induction, we find that $u_{k+1} \in W_0^{1,6r/(6m-(2m-3)r)}(\Omega)^3$, for each positive integer $2 \leqq m \leqq (3r-6)/(2r-6)$. By repeating the same argument, we arrive at $u_{k+1} \in W_0^{1,r}(\Omega)^3$.

Next we substitute u_{k+1} for w in (2.13) and derive

$$\varepsilon \sum_{k=0}^{N-1} \|u_{k+1}\|_V^2 \leqq M, \tag{2.15}$$

$$\|W_N\|_{C([0,T];L^2(\Omega)^3)} \leqq M, \tag{2.16}$$

$$\sum_{k=0}^{N-1} \|u_{k+1} - u_k\|_{L^2(\Omega)^3}^2 \leqq M, \tag{2.17}$$

where M denotes positive constants that are independent of η and N, and that depend only on $\|u_0\|_{L^2(\Omega)^3}$ and $\|f\|_{L^2(0,T;V')}$.

We now proceed to obtain the L^r estimates. Recalling that $0 \leqq \nu \leqq 3/4r$, we find that $u_{k+1} \in \mathcal{D}(A^{\nu})$, for $k = 0, 1, \cdots, N-1$, since $u_{k+1} \in V \cap W_0^{1,r}(\Omega)^3$. By making use of (1.26), (1.28), and (1.39), we can write (2.14) as

$$\begin{aligned} u_{k+1} &= (I + \varepsilon A)^{-1} u_k - \varepsilon A^{1/2+3/2r-2\nu}(I + \varepsilon A)^{-1} A^{-1/2-3/2r+2\nu} P \sum_{j=1}^{3} \partial_j (u_{kj} u_{k+1}) \\ &\quad + \varepsilon A^{1/2}(I + \varepsilon A)^{-1} A^{-1/2} P g \sum_{j=1}^{3} \partial_j \{(\eta + D_{\Pi}(u_{k+1}))^{-1/2} D_{ij}(u_{k+1})\} \\ &\quad + \varepsilon A^{1/2}(I + \varepsilon A)^{-1} A^{-1/2} P f_{k+1}, \end{aligned} \tag{2.18}$$

from which it follows that

$$\begin{aligned} u_k &= (I + \varepsilon A)^{-k} u_0 - \varepsilon \sum_{m=1}^{k} A^{1/2+3/2r-2\nu}(I + \varepsilon A)^{-(k-m+1)} A^{-1/2-3/2r+2\nu} P \sum_{j=1}^{3} \partial_j (u_{(m-1)j} u_m) \\ &\quad + \varepsilon \sum_{m=1}^{k} A^{1/2}(I + \varepsilon A)^{-(k-m+1)} A^{-1/2} P g \sum_{j=1}^{3} \partial_j \{(\eta + D_{\Pi}(u_m))^{-1/2} D_{ij}(u_m)\} \\ &\quad + \varepsilon \sum_{m=1}^{k} A^{1/2}(I + \varepsilon A)^{-(k-m+1)} A^{-1/2} P f_m \quad \text{for } k = 1, \cdots, N. \end{aligned} \tag{2.19}$$

Now we set

$$E_k = \|A^\nu u_k\|_{X_r} \quad \text{for } k = 0, 1, \cdots, N. \tag{2.20}$$

By virtue of (1.2), (1.7), (1.20), (1.22), and (1.39), we derive from (2.19)

$$\begin{aligned} E_k \leqq M_1 + \varepsilon M_2 \sum_{m=1}^{k} (\varepsilon(k-m+1))^{-(1/2+3/2r-\nu)} E_{m-1} E_m \\ + \varepsilon M_3 \sum_{m=1}^{k} (\varepsilon(k-m+1))^{-(1/2+\nu)} \quad \text{for } k = 1, \cdots, N, \end{aligned} \tag{2.21}$$

where M_i's are positive constants independent of η, k, and N. To estimate E_k from (2.21), we observe that

$$\varepsilon \sum_{m=1}^{k} (\varepsilon(k-m+1))^{-(1/2+3/2r-\nu)} \leqq (\varepsilon k)^{1/2-3/2r+\nu} \Big/ \left(\frac{1}{2} - \frac{3}{2r} + \nu\right), \tag{2.22}$$

$$\varepsilon \sum_{m=1}^{k} (\varepsilon(k-m+1))^{-(1/2+\nu)} \leqq (\varepsilon k)^{1/2-\nu} \Big/ \left(\frac{1}{2} - \nu\right). \tag{2.23}$$

We may assume that

$$M_1 \geqq \max(1, E_0) \tag{2.24}$$

and

$$0 < \varepsilon < 1 \quad \text{and} \quad 6M_1 M_2 \varepsilon^{1/2-3/2r+\nu} < \tfrac{1}{2}, \tag{2.25}$$

since we are interested only in large N.

We choose T_1 such that

$$0 < T_1 \leqq T, \tag{2.26}$$

$$M_3 T_1^{1/2-\nu} / (\tfrac{1}{2} - \nu) \leqq \tfrac{1}{2} M_1, \tag{2.27}$$

$$36 M_1 M_2 T_1^{1/2-3/2r+\nu} \Big/ \left(\frac{1}{2} - \frac{3}{2r} + \nu\right) \leqq 1. \tag{2.28}$$

Consequently, we find that if $\varepsilon = T/N < T_1$,

$$E_k \leqq 6M_1 \quad \text{for all } k \text{ such that } \varepsilon(k-1) \leqq T_1, \tag{2.29}$$

from which it follows that

$$\|A^\nu W_N\|_{C([0,T_1];X_r)} \leqq 6M_1, \tag{2.30}$$

for all large N such that $\varepsilon = T/N$ satisfies (2.25) and $\varepsilon < T_1$.

Next we obtain more regular estimates on W_N. Let us fix any ρ such that $0 < 2\rho < T_1$ and choose any s_1 and s_2 such that $\rho \leqq s_1 < s_2 \leqq T_1$. We then take N so large that (2.25) holds and

$$\varepsilon = \frac{T}{N} \leqq \frac{\rho}{2}. \tag{2.31}$$

Then there are positive integers k and n such that

$$k\varepsilon \leqq s_1 < (k+1)\varepsilon, \tag{2.32}$$

$$n\varepsilon \leqq s_2 < (n+1)\varepsilon. \tag{2.33}$$

We now choose any α and β such that

$$0<\alpha<\alpha+\beta<\frac{1}{2}-\frac{3}{2r}+\nu. \tag{2.34}$$

Then $u_m\in\mathscr{D}(A^{\nu+\alpha})$, for $m=1,\cdots,N$, since $u_m\in V\cap W_0^{1,r}(\Omega)^3$, for $m=1,\cdots,N$. To estimate $\|A^{\nu+\alpha}(W_N(s_2)-W_N(s_1))\|_{X_r}$, we treat three different cases separately.

In estimates (2.35)–(2.46), M denotes positive constants independent of η, ε, k, n, s_1, and s_2.

Case 1. $n-k\geqq 2$. According to (2.19), we write

$$\begin{aligned}
A^{\nu+\alpha}(u_n-u_k)=&A^{\alpha}((I+\varepsilon A)^{-n}-(I+\varepsilon A)^{-k})A^{\nu}u_0\\
&-\varepsilon\sum_{m=1}^{k}A^{\alpha+\beta+1/2+3/2r-\nu}((I+\varepsilon A)^{-(n-m+1)}\\
&-(I+\varepsilon A)^{-(k-m+1)})A^{-\beta}A^{-1/2-3/2r+2\nu}P\sum_{j=1}^{3}\partial_j(u_{(m-1)j}u_m)\\
&-\varepsilon\sum_{m=k+1}^{n}A^{\alpha+1/2+3/2r-\nu}\\
&\cdot(I+\varepsilon A)^{-(n-m+1)}A^{-1/2-3/2r+2\nu}P\sum_{j=1}^{3}\partial_j(u_{(m-1)j}u_m)\\
&+\varepsilon\sum_{m=1}^{k}A^{\alpha+\nu+1/2+\beta}((I+\varepsilon A)^{-(n-m+1)}\\
&-(I+\varepsilon A)^{-(k-m+1)})A^{-1/2-\beta}P\\
&\cdot\left(f_m+g\sum_{j=1}^{3}\partial_j\{(\eta+D_\Pi(u_m))^{-1/2}D_{ij}(u_m)\}\right)\\
&+\varepsilon\sum_{m=k+1}^{n}A^{\alpha+\nu+1/2}(I+\varepsilon A)^{-(n-m+1)}A^{-1/2}P\\
&\cdot\left(f_m+g\sum_{j=1}^{3}\partial_j\{(\eta+D_\Pi(u_m))^{-1/2}D_{ij}(u_m)\}\right)
\end{aligned} \tag{2.35}$$

and estimate each term on the right-hand side.

$$\begin{aligned}
&\|A^{\alpha}((I+\varepsilon A)^{-n}-(I-\varepsilon A)^{-k})A^{\nu}u_0\|_{X_r}\\
&\quad\leqq\|A(I+\varepsilon A)^{-k}((I+\varepsilon A)^{-(n-k)}-I)A^{-(1-\alpha)}A^{\nu}u_0\|_{X_r}\\
&\quad\leqq\frac{M}{\varepsilon k}(\varepsilon(n-k))^{1-\alpha}\|A^{\nu}u_0\|_{X_r}\leqq\frac{M}{\rho}(s_2-s_1)^{1-\alpha}\quad\text{using (1.2) and (1.9),}
\end{aligned} \tag{2.36}$$

$$\begin{aligned}
&\varepsilon\sum_{m=1}^{k}\|A^{\alpha+\beta+1/2+3/2r-\nu}(I+\varepsilon A)^{-(k-m+1)}((I+\varepsilon A)^{-(n-k)}-I)A^{-\beta}\\
&\qquad\cdot A^{-1/2-3/2r+2\nu}P\sum_{j=1}^{3}\partial_j(u_{(m-1)j}u_m)\|_{X_r}\\
&\quad\leqq M(\varepsilon(n-k))^{\beta}\sum_{m=1}^{k}\varepsilon(\varepsilon(k-m+1))^{-(\alpha+\beta+1/2+3/2r-\nu)}E_{m-1}E_m,\\
&\qquad\qquad\text{using (1.2), (1.9), (1.22), and (2.20),}\\
&\quad\leqq M(\varepsilon(n-k))^{\beta}(\varepsilon k)^{1-(\alpha+\beta+1/2+3/2r-\nu)}/\{1-(\alpha+\beta+\tfrac{1}{2}+3/2r-\nu)\}\\
&\quad\leqq M(s_2-s_1)^{\beta}T_1^{1-(\alpha+\beta+1/2+3/2r-\nu)}\quad\text{using (2.29),}
\end{aligned} \tag{2.37}$$

$$\varepsilon \sum_{m=k+1}^{n} \|A^{\alpha+1/2+3/2r-\nu}(I+\varepsilon A)^{-(n-m+1)}A^{-1/2-3/2r+2\nu}P \sum_{j=1}^{3} \partial_j(u_{(m-1)j}u_m)\|_{X_r}$$

$$\leqq M \sum_{m=k+1}^{n} \varepsilon(\varepsilon(n-m+1))^{-(\alpha+1/2+3/2r-\nu)}E_{m-1}E_m \tag{2.38}$$

$$\leqq M(s_2-s_1)^{1/2-\alpha-3/2r+\nu},$$

$$\varepsilon \sum_{m=1}^{k} \|A^{\alpha+\nu+1/2+\beta}((I+\varepsilon A)^{-(n-m+1)}-(I+\varepsilon A)^{-(k-m+1)})A^{-\beta}A^{-1/2}P$$

$$\cdot\left(f_m+g\sum_{j=1}^{3}\partial_j\{(\eta+D_\Pi(u_m))^{-1/2}D_{ij}(u_m)\}\right)\|_{X_r} \tag{2.39}$$

$$\leqq M(\varepsilon(n-k))^{\beta}\sum_{m=1}^{k}\varepsilon(\varepsilon(k-m+1))^{-(\alpha+\nu+1/2+\beta)}$$

$$\leqq M(s_2-s_1)^{\beta}T_1^{1/2-(\alpha+\nu+\beta)} \quad \text{using (1.2), (1.9), and (1.39),}$$

$$\varepsilon \sum_{m=k+1}^{n} \left\| A^{\alpha+\nu+1/2}(I+\varepsilon A)^{-(n-m+1)}A^{-1/2}P \right.$$

$$\left. \cdot\left(f_m+g\sum_{j=1}^{3}\partial_j\{(\eta+D_\Pi(u_m))^{-1/2}D_{ij}(u_m)\}\right)\right\|_{X_r}$$

$$\leqq M\sum_{m=k+1}^{n}\varepsilon(\varepsilon(n-m+1))^{-(\alpha+\nu+1/2)} \tag{2.40}$$

$$\leqq M(\varepsilon(n-k))^{1-(\alpha+\nu+1/2)}$$

$$\leqq M(s_2-s_1)^{1/2-\alpha-\nu}.$$

Combining the above estimates, we have

$$\|A^{\alpha+\nu}(u_n-u_k)\|_{X_r} \leqq \frac{M}{\rho}(s_2-s_1)^{1-\alpha}+MT_1^{1/2+\nu-\alpha-\beta-3/2r}(s_2-s_1)^{\beta}$$

$$+M(s_2-s_1)^{1/2+\nu-\alpha-3/2r}+MT_1^{1/2-\alpha-\nu-\beta}(s_2-s_1)^{\beta} \tag{2.41}$$

$$+M(s_2-s_1)^{1/2-\alpha-\nu}$$

$$\leqq M(s_2-s_1)^{\beta}.$$

Similarly, we can also obtain

$$\|A^{\alpha+\nu}(u_{n+1}-u_{k+1})\|_{X_r} \leqq M(s_2-s_1)^{\beta}, \tag{2.42}$$

$$\|A^{\alpha+\nu}(u_{k+1}-u_k)\|_{X_r} \leqq M\varepsilon^{\beta}. \tag{2.43}$$

Since $s_1=k\varepsilon+\lambda_1\varepsilon$, $s_2=n\varepsilon+\lambda_2\varepsilon$ with $0\leqq\lambda_1<1$, $0\leqq\lambda_2<1$, and $s_2-s_1>\varepsilon$,

$$\|A^{\alpha+\nu}(W_N(s_2)-W_N(s_1))\|_{X_r} \leqq (1-\lambda_2)\|A^{\alpha+\nu}(u_n-u_k)\|_{X_r}$$

$$+\lambda_2\|A^{\alpha+\nu}(u_{n+1}-u_{k+1})\|_{X_r}+|\lambda_2-\lambda_1|\,\|A^{\alpha+\nu}(u_{k+1}-u_k)\|_{X_r} \tag{2.44}$$

$$\leqq M(s_2-s_1)^{\beta}.$$

Case 2. $n=k+1$. In this case, $s_1=(k+1)\varepsilon-\lambda_1\varepsilon$ and $s_2=(k+1)\varepsilon+\lambda_2\varepsilon$, where $0\leqq\lambda_2<1$, $0<\lambda_1\leqq 1$ and $s_2-s_1=(\lambda_1+\lambda_2)\varepsilon$. Hence, (2.43) yields

$$
\begin{aligned}
\|A^{\alpha+\nu}(W_N(s_2)-W_N(s_1))\|_{X_r} &\leqq \|A^{\alpha+\nu}((1-\lambda_2)u_{k+1}+\lambda_2 u_{k+2}) \\
&\quad -A^{\alpha+\nu}(\lambda_1 u_k+(1-\lambda_1)u_{k+1})\|_{X_r} \\
&\leqq \lambda_1\|A^{\alpha+\nu}(u_{k+1}-u_k)\|_{X_r}+\lambda_2\|A^{\alpha+\nu}(u_{k+2}-u_{k+1})\|_{X_r} \\
&\leqq (\lambda_1+\lambda_2)M\varepsilon^{\beta}=M\varepsilon^{\beta}\frac{s_2-s_1}{\varepsilon}=M(s_2-s_1)^{\beta}\left(\frac{s_2-s_1}{\varepsilon}\right)^{1-\beta} \\
&\leqq 2^{1-\beta}M(s_2-s_1)^{\beta}\leqq M(s_2-s_1)^{\beta}.
\end{aligned}
\tag{2.45}
$$

Case 3. $n=k$. Using $0<s_2-s_1<\varepsilon$ and (2.43), we have

$$
\begin{aligned}
\|A^{\alpha+\nu}(W_N(s_2)-W_N(s_1))\|_{X_r} &= \frac{s_2-s_1}{\varepsilon}\|A^{\alpha+\nu}(u_{k+1}-u_k)\|_{X_r} \\
&\leqq M(s_2-s_1)^{\beta}\left(\frac{s_2-s_1}{\varepsilon}\right)^{1-\beta} \\
&\leqq M(s_2-s_1)^{\beta}.
\end{aligned}
\tag{2.46}
$$

On account of (2.44)-(2.46), we conclude that if $\varepsilon=T/N$ satisfies (2.25) and (2.31), then

$$
\|A^{\alpha+\nu}(W_N(s_2)-W_N(s_1))\|_{X_r}\leqq M(s_2-s_1)^{\beta} \quad \text{for } \rho\leqq s_1<s_2\leqq T_1. \tag{2.47}
$$

Through an analogous procedure, we can also derive

$$
\|A^{\alpha+\nu}W_N(s)\|_{X_r}\leqq M \quad \text{for } \rho\leqq s\leqq T_1. \tag{2.48}
$$

In (2.47) and (2.48), M stands for positive constants independent of η, ε, s_1, and s_2, and dependent on α, β, ν, ρ, T_1, $\|A^{\nu}u_0\|_{X_r}$, and $\|f\|_{L^{\infty}(0,T;W^{-1,r}(\Omega)^3)}$.

2.2. Convergence as the time meshsize tends to zero. In this section, we prove that the approximate solution W_N defined by (2.9) converges to a solution of the regularized problem as $N\to\infty$. We recall that $0\leqq\nu\leqq 3/4r$, $u_0(x)\in\mathscr{D}(A^{\nu})$, and $f\in L^{\infty}(0,T;W^{-1,r}(\Omega)^3)$, and that T_1 was chosen according to (2.26)-(2.28).

LEMMA 2.5. *The sequence* $\{W_N\}_{N=1}^{\infty}$ *is precompact in* $C([0,T_1];\mathscr{D}(A^{\nu}))$.

Proof. First we fix a positive integer N_0 such that $T/N_0<T_1$ and $\varepsilon=T/N_0$ satisfies (2.25). Let $t^*\in(0,T_1]$ be given. Then, we can choose an integer $N^*\geqq N_0$ and a positive number ρ so that $2\rho<t^*\leqq T_1$ and $\varepsilon=T/N\leqq\rho/2$, for all $N\geqq N^*$. By virtue of (2.48), it is evident that $\{W_N(t^*)\}_{N=N^*}^{\infty}$ is precompact in $\mathscr{D}(A^{\nu})$ and, consequently, $\{W_N(t^*)\}_{N=N_0}^{\infty}$ is precompact in $\mathscr{D}(A^{\nu})$. Furthermore, it follows from (2.47) that $\{W_N\}_{N=N^*}^{\infty}$ is equicontinuous at $t=t^*$ where each W_N is regarded as a continuous function from $[0,T_1]$ into $\mathscr{D}(A^{\nu})$. Thus, $\{W_N\}_{N=N_0}^{\infty}$ is equicontinuous at t^*. Next we show that $\{W_N\}_{N=N_0}^{\infty}$ is equicontinuous at $t=0$. Let us recall that $W_N(0)=u_0$ for every N, and derive an analogue of (2.21) from (2.19):

$$
\begin{aligned}
\|A^{\nu}(u_k-u_0)\|_{X_r} &\leqq \|((I+\varepsilon A)^{-k}-I)A^{\nu}u_0\|_{X_r} \\
&\quad +\varepsilon M_2\sum_{m=1}^{k}(\varepsilon(k-m+1))^{-(1/2+3/2r-\nu)}E_{m-1}E_m \\
&\quad +\varepsilon M_3\sum_{m=1}^{k}(\varepsilon(k-m+1))^{-(1/2+\nu)} \quad \text{for } k=1,\cdots,N,
\end{aligned}
\tag{2.49}
$$

where M_2 and M_3 are the same as in (2.21). Let any $\zeta > 0$ be given. By assuming $N \geqq N_0$ and combining Lemma 1.4, (2.22), (2.23), (2.29), and (2.49), we can find a positive number δ depending on ζ and u_0, but independent of η, N, and k such that $\varepsilon k \leqq \delta$ implies

$$\|A^{\nu}(u_k - u_0)\|_{X_r} \leqq \zeta. \tag{2.50}$$

Using this δ, we set

$$\tilde{\delta} = \min\left(\frac{\delta}{2}, \frac{\delta}{2}\frac{\zeta}{7M_1}\right), \tag{2.51}$$

where M_1 is the same as in (2.29); see also (2.24).

If $N \geqq N_0$, $N \geqq 2T/\delta$, and $0 \leqq s \leqq \tilde{\delta}$, then (2.50) gives

$$\|A^{\nu}(W_N(s) - u_0)\|_{X_r} \leqq \zeta. \tag{2.52}$$

If $N \geqq N_0$, $N < 2T/\delta$, and $0 \leqq s \leqq \tilde{\delta}$, then $\varepsilon = T/N > \delta/2 \geqq \tilde{\delta}$ and

$$\begin{aligned} \|A^{\nu}(W_N(s) - u_0)\|_{X_r} &\leqq \frac{s}{\varepsilon}\|A^{\nu}(u_1 - u_0)\|_{X_r} \\ &\leqq \frac{s}{\varepsilon} 7M_1 \quad \text{by (2.24) and (2.29),} \\ &\leqq \zeta \quad \text{by (2.51).} \end{aligned} \tag{2.53}$$

We now conclude that for any given $\zeta > 0$, there is $\tilde{\delta} > 0$ independent of η and N such that

$$\|A^{\nu}(W_N(s) - u_0)\|_{X_r} \leqq \zeta \quad \text{for all } s \in [0, \tilde{\delta}] \text{ and all } N \geqq N_0. \tag{2.54}$$

Hence, $\{W_N\}_{N=N_0}^{\infty}$ is equicontinuous at $t = 0$. According to the Ascoli Theorem, $\{W_N\}_{N=N_0}^{\infty}$ is precompact in $C([0, T_1]; \mathcal{D}(A^{\nu}))$ and so is $\{W_N\}_{N=1}^{\infty}$. This completes the proof of Lemma 2.5. □

We can now extract a subsequence still denoted by $\{W_N\}$ such that for some function u

$$\lim_{N\to\infty} W_N = u \tag{2.55}$$

in the norm of $C([0, T_1]; \mathcal{D}(A^{\nu}))$.

LEMMA 2.6. *The limit u in* (2.55) *satisfies*

$$u \in L^2(0, T_1; V), \qquad \partial_t u \in L^2(0, T_1; V'), \tag{2.56}$$

$$u(x, 0) = u_0(x), \tag{2.57}$$

$$(\partial_t u, w - u) + a(u, w - u) + b(u, u, w) + J_{\eta}(w) - J_{\eta}(u) \geqq (f, w - u) \tag{2.58}$$

for all $w \in V$, for almost all $t \in (0, T_1)$. Furthermore, it holds that for any $0 < \rho < T_1$ and α, β satisfying (2.34),

$$\|u(s)\|_{\mathcal{D}(A^{\alpha+\nu})} \leqq M \quad \text{for } \rho \leqq s \leqq T_1, \tag{2.59}$$

$$\|u(s_2) - u(s_1)\|_{\mathcal{D}(A^{\alpha+\nu})} \leqq M(s_2 - s_1)^{\beta} \quad \text{for } \rho \leqq s_1 < s_2 \leqq T_1, \tag{2.60}$$

where M denotes positive constants independent of η, s, s_1, and s_2, and dependent on α, β, ν, ρ, T_1, $\|u_0\|_{\mathcal{D}(A^{\nu})}$, and $\|f\|_{L^{\infty}(0,T;W^{-1,r}(\Omega)^3)}$.

Proof. Equation (2.57) is obvious since $W_N(0)=u_0$, for each N. By virtue of (2.47), (2.48), and Lemma 1.1, we can use the Ascoli Theorem to obtain (2.59) and (2.60). Next let us fix a positive integer N_0 such that $T/N_0<T_1$ and $\varepsilon=T/N_0$ satisfies (2.25).

Then, it follows from (1.39), (2.14), (2.15), and (2.29) that for $N\geqq N_0$ and $T/N\leqq\rho<T_1$,

$$\|W_N\|_{L^2(\rho,T_1;V)}\leqq M, \tag{2.61}$$

$$\|\partial_t W_N\|_{L^2(\rho,T_1;V')}\leqq M, \tag{2.62}$$

where M denotes positive constants independent of η, ρ, and N. Consequently, (2.55) implies that for each $0<\rho<T_1$,

$$W_N\to u \quad \text{weakly in } L^2(\rho, T_1; V) \quad \text{as } N\to\infty, \tag{2.63}$$

$$\partial_t W_N\to\partial_t u \quad \text{weakly in } L^2(\rho, T_1; V') \quad \text{as } N\to\infty, \tag{2.64}$$

and consequently,

$$\|u\|_{L^2(0,T_1;V)}\leqq M, \tag{2.65}$$

$$\|\partial_t u\|_{L^2(0,T_1;V')}\leqq M, \tag{2.66}$$

where M stands for positive constants independent of η.

We proceed to prove (2.58).

Let us define

$$U_N(t)=u_{k+1} \quad \text{for } k\varepsilon\leqq t<(k+1)\varepsilon, \qquad k=0,1,\cdots,N-1, \tag{2.67}$$

$$U_N^*(t)=u_k \quad \text{for } k\varepsilon\leqq t<(k+1)\varepsilon, \qquad k=0,1,\cdots,N-1, \tag{2.68}$$

$$F_N(t)=f_{k+1} \quad \text{for } k\varepsilon\leqq t<(k+1)\varepsilon, \qquad k=0,1,\cdots,N-1. \tag{2.69}$$

Then it is proved in [15, p. 329] that as $N\to\infty$,

$$F_N\to f \quad \text{strongly in } L^2(0, T; V'), \tag{2.70}$$

$$W_N-U_N\to 0 \quad \text{strongly in } L^2(0, T; X_2). \tag{2.71}$$

Likewise, by using (2.17), it is easy to see that

$$W_N-U_N^*\to 0 \quad \text{strongly in } L^2(0, T; X_2). \tag{2.72}$$

Furthermore, it follows from (2.15) and (2.29) that

$$\|U_N\|_{L^2(0,T;V)}\leqq M \quad \text{for all } N, \tag{2.73}$$

$$\|W_N\|_{L^2(\rho,T;V)}\leqq M \quad \text{for all } \rho,\ N \text{ such that } T/N\leqq\rho<T, \tag{2.74}$$

$$\|U_N^*\|_{L^\infty(0,T_1;X_r)}\leqq M \quad \text{for all } N\geqq N_0, \tag{2.75}$$

where M stands for positive constants independent of η, ρ, and N.

Consequently, (2.55) implies that as $N\to\infty$

$$U_N\to u \quad \text{strongly in } L^2(0, T_1; X_2), \tag{2.76}$$

$$U_N^*\to u \quad \text{strongly in } L^2(0, T_1; X_2), \tag{2.77}$$

$$U_N\to u \quad \text{weakly in } L^2(0, T_1; V), \tag{2.78}$$

which, together with (2.73) and (2.75), give

$$U_{Nj}^*U_N\to u_j u \quad \text{weakly in } L^2(0, T_1; L^2(\Omega)^3) \tag{2.79}$$

for $j=1,2,3$, where $U_N^*=(U_{N1}^*, U_{N2}^*, U_{N3}^*)$ and $u=(u_1, u_2, u_3)$.

Next it follows from (2.10) that for each N,

$$\int_0^{T_1} (\partial_t W_N, \psi)\, dt + \int_0^{T_1} a(U_N, \psi)\, dt + \int_0^{T_1} b(U_N^*, U_N, \psi)\, dt + \int_0^{T_1} (J_\eta'(U_N), \psi)\, dt = \int_0^{T_1} (F_N, \psi)\, dt, \tag{2.80}$$

for all $\psi \in L^2(0, T_1; V)$. By making use of (1.38) and the fact that $U_N \in L^2(0, T; V)$, we infer that for each N,

$$\begin{aligned} &\int_0^{T_1} (\partial_t W_N, \psi - U_N)\, dt + \int_0^{T_1} a(U_N, \psi - U_N)\, dt \\ &\quad + \int_0^{T_1} b(U_N^*, U_N, \psi)\, dt + \int_0^{T_1} (J_\eta(\psi) - J_\eta(U_N))\, dt \\ &\quad \geqq \int_0^{T_1} (F_N, \psi - U_N)\, dt \end{aligned} \tag{2.81}$$

for all $\psi \in L^2(0, T_1; V)$. We next observe that for $k\varepsilon \leqq t < (k+1)\varepsilon$, $k = 0, 1, \cdots, N-1$,

$$(\partial_t W_N, U_N - W_N) = \frac{1}{\varepsilon}\left(1 - \frac{1}{\varepsilon}(t - k\varepsilon)\right) \|u_{k+1} - u_k\|_{X_2}^2 \geqq 0, \tag{2.82}$$

and hence,

$$\begin{aligned} \int_0^{T_1} (\partial_t W_N, U_N)\, dt &= \int_0^{T_1} (\partial_t W_N, W_N)\, dt + \int_0^{T_1} (\partial_t W_N, U_N - W_N)\, dt \\ &\geqq \int_0^{T_1} (\partial_t W_N, W_N)\, dt \\ &= \frac{1}{2} \|W_N(T_1)\|_{X_2}^2 - \frac{1}{2} \|u_0\|_{X_2}^2. \end{aligned} \tag{2.83}$$

Accordingly, we use (2.55), (2.65), and (2.66) to deduce

$$\begin{aligned} \liminf_{N\to\infty} \int_0^{T_1} (\partial_t W_N, U_N)\, dt &\geqq \liminf_{N\to\infty} \int_0^{T_1} (\partial_t W_N, W_N)\, dt \\ &= \frac{1}{2} \|u(T_1)\|_{X_2}^2 - \frac{1}{2} \|u_0\|_{X_2}^2 \\ &= \int_0^{T_1} (\partial_t u, u)\, dt. \end{aligned} \tag{2.84}$$

By virtue of (2.78) and (2.79), we see that

$$\liminf_{N\to\infty} \int_0^{T_1} J_\eta(U_N)\, dt \geqq \int_0^{T_1} J_\eta(u)\, dt, \tag{2.85}$$

$$\liminf_{N\to\infty} \int_0^{T_1} a(U_N, U_N)\, dt \geqq \int_0^{T_1} a(u, u)\, dt, \tag{2.86}$$

$$\lim_{N\to\infty} \int_0^{T_1} b(U_N^*, U_N, \psi)\, dt = \int_0^{T_1} b(u, u, \psi)\, dt \quad \text{for each } \psi \in L^2(0, T_1; V). \tag{2.87}$$

Combining (2.64), (2.70), and (2.84)-(2.87), we can pass $N \to \infty$ in (2.81) to arrive at

$$\begin{aligned} &\int_0^{T_1} (\partial_t u, \psi - u)\, dt + \int_0^{T_1} a(u, \psi - u)\, dt + \int_0^{T_1} b(u, u, \psi)\, dt \\ &\quad + \int_0^{T_1} (J_\eta(v) - J_\eta(u))\, dt \\ &\quad \geqq \int_0^{T_1} (f, \psi - u)\, dt \end{aligned} \tag{2.88}$$

for every $\psi \in L^2(0, T_1; V)$. By the same argument as in [4], (2.88) implies (2.58). This ends the proof of Lemma 2.6. □

2.3. Convergence as η tends to zero. For each $\eta > 0$, we denote by u_η the function in Lemma 2.6 to signify its dependence on η. We will show that u_η converges to a solution of (0.1)-(0.4) as η tends to zero. Since T_1 was chosen according to (2.26)-(2.28), T_1 is independent of η and nonincreasing in $\|u_0\|_{\mathcal{D}(A^\nu)}$.

LEMMA 2.7. *The set $\{u_\eta\}_{\eta>0}$ is precompact in $C([0, T_1]; \mathcal{D}(A^\nu))$.*

Proof. By (2.54) and (2.55), it is easily seen that $\{u_\eta(t)\}_{\eta>0}$ is equicontinuous at $t = 0$. Inequalities (2.59) and (2.60) imply that $\{u_\eta(t)\}_{\eta>0}$ is precompact in $\mathcal{D}(A^\nu)$ and equicontinuous at each $0 < t \leqq T_1$. Now we apply the Ascoli theorem to conclude the proof. □

Let us extract a sequence still denoted by $\{u_\eta\}$ such that for some function u,

$$\lim_{\eta \to 0} u_\eta = u \tag{2.89}$$

in the norm of $C([0, T_1]; \mathcal{D}(A^\nu))$.

LEMMA 2.8. *The limit function u in* (2.89) *is a solution of* (0.1)-(0.4) *on the interval* $[0, T_1)$. *Furthermore, it holds that for any $0 < \rho < T_1$ and α, β satisfying* (2.34),

$$\|u(s)\|_{\mathcal{D}(A^{\alpha+\nu})} \leqq M \quad \text{for } \rho \leqq s \leqq T_1, \tag{2.90}$$

$$\|u(s_2) - u(s_1)\|_{\mathcal{D}(A^{\alpha+\nu})} \leqq M(s_2 - s_1)^\beta \quad \text{for } \rho \leqq s_1 < s_2 \leqq T_1, \tag{2.91}$$

where M denotes positive constants independent of s, s_1, and s_2, and dependent on α, β, ν, ρ, T_1, $\|u_0\|_{\mathcal{D}(A^\nu)}$, and $\|f\|_{L^\infty(0,T;W^{-1,r}(\Omega)^3)}$.

Proof. Since the positive constants denoted by M in (2.59) and (2.60) are independent of η, we use Lemma 1.1 and the Ascoli Theorem to derive (2.90) and (2.91). Meanwhile, (2.65), (2.66), and (2.89) yield

$$u_\eta \to u \quad \text{weakly in } L^2(0, T_1; V), \tag{2.92}$$

$$\partial_t u_\eta \to \partial_t u \quad \text{weakly in } L^2(0, T_1; V'), \tag{2.93}$$

and thus

$$\begin{aligned} \liminf_{\eta \to 0} \int_0^{T_1} J_\eta(u_\eta)\, dt &\geqq \liminf_{\eta \to 0} \int_0^{T_1} J(u_\eta)\, dt \\ &\geqq \int_0^{T_1} J(u)\, dt. \end{aligned} \tag{2.94}$$

Now, by the same argument as in the proof of Lemma 2.6, we can show that u satisfies (0.1) for every $w \in V$, for almost all $t \in (0, T_1)$. We omit the details.

2.4. Conclusion of the proof of Theorem 2.3. It remains to establish the uniqueness of solution. Let u be the solution obtained in Lemma 2.8, and let $\tilde{u}$ be another solution on $[0, T_1)$ according to Definition 2.1. Then, we have

$$(\partial_t(u-\tilde{u}), u-\tilde{u}) + a(u-\tilde{u}, u-\tilde{u}) \leqq b(u, u, \tilde{u}) + b(\tilde{u}, \tilde{u}, u), \tag{2.95}$$

for almost all $t \in (0, T_1)$.

As in [11, p. 85], we have

$$\begin{aligned} |b(u, u, \tilde{u}) + b(\tilde{u}, \tilde{u}, u)| &= |b(u-\tilde{u}, u-\tilde{u}, u)| \\ &\leqq M\|u\|_{X_r}\|u-\tilde{u}\|_{X_2}^{2/s}\|u-\tilde{u}\|_V^{1+3/r}, \end{aligned} \tag{2.96}$$

where $2/s + 3/r = 1$ and M is a positive constant. Since $u \in C([0, T_1]; X_r)$, we combine (2.95), (2.96), and the Hölder inequality to deduce

$$(\partial_t(u-\tilde{u}), u-\tilde{u}) \leqq M\|u-\tilde{u}\|_{X_2}^2, \tag{2.97}$$

for almost all $t \in (0, T_1)$.

It now follows that $u = \tilde{u}$ on $[0, T_1]$.

This and Lemma 2.8 complete the proof of Theorem 2.3. □

Finally, we remark that if $T_1 < T$, then the solution can be extended to a larger interval by means of Theorem 2.3 itself. Let us choose

$$T^* = \sup\{\tilde{T}: T_1 \leqq \tilde{T} \leqq T \text{ and there is a solution that is } \mathscr{D}(A^\nu)\text{-valued continuous on } [0, \tilde{T}]\}. \tag{2.98}$$

Then, $u \in C([0, T_1]; \mathscr{D}(A^\nu))$ can be extended to $u \in C([0, T^*); \mathscr{D}(A^\nu))$. If $T^* < T$,

$$\lim_{t \to T^*-} \|u(t)\|_{\mathscr{D}(A^\nu)} = \infty. \tag{2.99}$$

In this case, it is not known whether there is a solution (according to Definition 2.1) defined on an interval $[0, T^{**})$, $T^{**} > T^*$.

3. Global existence. Our assertion on the existence of global solution is Theorem 3.1.

THEOREM 3.1. *Let*

$$3 < \frac{6r}{r+3} < p < r \quad \text{and} \quad \lambda = \frac{r-(2r/p)}{r-2}.$$

There is a positive number C such that if $u_0 \in X_r$ and $f \in L^\infty(0, \infty; W^{-1,r}(\Omega)^3)$ satisfy

$$(\|u_0\|_{X_2} + \|f\|_{L^\infty(0,\infty;V')})^{2(1-\lambda)}(1 + \|u_0\|_{X_r} + \|f\|_{L^\infty(0,\infty;W^{-1,r}(\Omega)^3)})^\lambda \leqq C, \tag{3.1}$$

then there is a unique solution u according to Definition 2.1 on the interval $[0, T)$, for any $0 < T < \infty$. Furthermore, $u \in C([0, \infty); X_r)$ and, for each $0 < \delta \leqq \frac{1}{2}$, u is $\mathscr{D}(A^{1/2-\delta})$-valued and locally Hölder continuous on $(0, \infty)$.

The idea of proof is to choose any $T > 0$ and establish the existence of solution u on $[0, T]$ together with the estimate of $\sup_{t \in [0,T]} \|u(t)\|_{X_r}$. Under assumption (3.1), it will be shown that this estimate is independent of T, and hence the time interval can be extended indefinitely with the aid of the uniqueness of solutions.

Proof. We will follow the scheme of § 2.1. Choose any $T > 0$ and any positive integer N. As before, we set $\varepsilon = T/N$. Let us write

$$G_k = \|u_k\|_{X_2} \quad \text{for } k = 0, 1, \cdots, N, \tag{3.2}$$

$$\Theta_1 = \operatorname{ess\,sup}_{t \geqq 0} \|f(t)\|_{V'}. \tag{3.3}$$

By substituting u_{k+1} for ϕ in (2.10), we can derive

$$G_{k+1}^2 - G_k^2 + \varepsilon C_1 G_{k+1}^2 \leqq \varepsilon C_2 \Theta_1^2 \quad \text{for } k = 0, 1, \cdots, N-1, \tag{3.4}$$

where C_1 and C_2 are positive constants dependent only on Ω and μ. Since we are interested only in small ε, we may assume

$$0 < \varepsilon < \min\left(1, \frac{1}{2C_1}\right). \tag{3.5}$$

It then follows from (3.4) that

$$\begin{aligned} G_k^2 &\leqq (1+\varepsilon C_1)^{-k} G_0^2 + \varepsilon C_2 \Theta_1^2 \sum_{m=1}^{k} (1+\varepsilon C_1)^{-m}, \\ &\leqq G_0^2 \exp(-\delta_1 \varepsilon k) + \frac{C_2}{C_1} \Theta_1^2 \quad \text{for } k = 1, \cdots, N, \end{aligned} \tag{3.6}$$

where δ_1 is a positive number independent of ε, k, and T.

We will obtain a new estimate of $E_k = \|u_k\|_{X_r}$: we are considering the case $\nu = 0$ in (2.20).

Recalling that $3 < 6r/(r+3) < p < r$, we set

$$b = \frac{1}{2} + \frac{3}{2}\left(\frac{2}{p} - \frac{1}{r}\right) \tag{3.7}$$

and rewrite (2.19) by using (1.27) as

$$\begin{aligned} u_k &= (I+\varepsilon A)^{-k} u_0 - \varepsilon \sum_{m=1}^{k} A^b (I+\varepsilon A)^{-(k-m+1)} A^{-b} P \sum_{j=1}^{3} \partial_j (u_{(m-1)j} u_m) \\ &+ \varepsilon \sum_{m=1}^{k} A^{1/2} (I+\varepsilon A)^{-(k-m+1)} A^{-1/2} P g \sum_{j=1}^{3} \partial_j \{(\eta + D_\Pi(u_m))^{-1/2} D_{ij}(u_m)\} \\ &+ \varepsilon \sum_{m=1}^{k} A^{1/2} (I+\varepsilon A)^{-(k-m+1)} A^{-1/2} P f_m \quad \text{for } k = 1, \cdots, N. \end{aligned} \tag{3.8}$$

By means of (1.2), (1.20), (1.21), (1.39), and the inequality

$$\|h\|_{L^p(\Omega)} \leqq \|h\|_{L^2(\Omega)}^{1-\lambda} \|h\|_{L^r(\Omega)}^{\lambda} \quad \text{for all } h \in L^r(\Omega) \quad \text{where } \lambda = \frac{r-(2r/p)}{r-2}, \tag{3.9}$$

we deduce

$$\begin{aligned} E_k &\leqq C_3 \exp(-\delta_2 \varepsilon k) E_0 + C_4 \varepsilon \sum_{m=1}^{k} \exp(-\delta_3 \varepsilon (k-m+1)) \\ &\quad \cdot (\varepsilon(k-m+1))^{-b} G_{m-1}^{1-\lambda} G_m^{1-\lambda} E_{m-1}^{\lambda} E_m^{\lambda} \\ &\quad + C_5 \varepsilon \sum_{m=1}^{k} \exp(-\delta_4 \varepsilon (k-m+1)) (\varepsilon(k-m+1))^{-1/2} (\Theta_2 + C_6), \end{aligned} \tag{3.10}$$

where $\Theta_2 = \operatorname{ess\,sup}_{t \geqq 0} \|f(t)\|_{W^{-1,r}(\Omega)^3}$, and C_i's and δ_i's are positive constants independent of η, ε, k, T, E_m's, G_m's, Θ_1, and Θ_2.

Let us set

$$C_7 = C_4 \int_0^\infty \exp(-\delta_3 s) s^{-b}\, ds, \tag{3.11}$$

$$C_8 = C_5 \int_0^\infty \exp(-\delta_4 s) s^{-1/2}\, ds, \tag{3.12}$$

$$C_9 = \max(1, C_3, C_8), \tag{3.13}$$

$$M_1 = \max(1, E_0, \Theta_2 + C_6), \tag{3.14}$$

$$M_2 = 4C_9 M_1. \tag{3.15}$$

We then choose $\zeta > 0$ such that

$$C_4 \zeta^{2(1-\lambda)} M_2^{\lambda} \leqq \tfrac{1}{3}, \tag{3.16}$$

$$C_7 \zeta^{2(1-\lambda)} M_2^{\lambda} \leqq \tfrac{1}{6}, \tag{3.17}$$

and suppose that

$$G_0 + \left(\frac{C_2}{C_1}\right)^{1/2} \Theta_1 \leqq \zeta. \tag{3.18}$$

Then, by virtue of (3.6),

$$G_k \leqq \zeta \quad \text{for all } k = 0, 1, \cdots, N. \tag{3.19}$$

We proceed to show by induction

$$E_k \leqq M_2 \quad \text{for all } k = 0, \cdots, N. \tag{3.20}$$

It is obvious that $E_0 \leqq M_2$ and $M_2 \geqq 4$.

If $E_m \leqq M_2$, for $m = 0, \cdots, k-1$, then by (3.10), (3.16), and (3.17), we have

$$E_k \leqq 2C_9 M_1 + \tfrac{1}{6} M_2 + \tfrac{1}{3} E_k^{\lambda}. \tag{3.21}$$

If $E_k \leqq 1$, then $E_k \leqq M_2$.

If $E_k > 1$, then (3.21) yields

$$\tfrac{2}{3} E_k \leqq 2C_9 M_1 + \tfrac{1}{6} M_2 \tag{3.22}$$

and hence $E_k \leqq M_2$. This proves (3.20).

Now we can choose a proper positive constant C in (3.1), subject to the constants above, independent of u_0 and f so that (3.1) guarantees (3.16)–(3.18), from which (3.20) follows. The remainder of the proof of Theorem 3.1 can be carried out precisely in the same manner as in the previous section, and we omit it. □

Next we present a result on the asymptotic behavior of solutions when the external force f is time-periodic. Our assertion is given in Theorem 3.2.

THEOREM 3.2. *Suppose that f is L-periodic in time and that u_0 and f satisfy* (3.1) *and*

$$(\|u_0\|_{X_2} + \Theta_1)^d (1 + \|u_0\|_{X_r} + \Theta_2)^{1-d} \leqq \tilde{C}, \tag{3.23}$$

where Θ_1, Θ_2 are the same as above, and d, $\tilde{C}$ are positive constants that will be determined below. Then the solution $u(x, t)$ of the theorem above converges to an L-periodic solution as $t \to \infty$. More precisely, there is a function $u_L(x, t)$ satisfying (0.1) *on* $(-\infty, \infty)$ *such that*

$$u_L \in L^2(0, L; V) \quad \text{and} \quad \partial_t u_L \in L^2(0, L; V'), \tag{3.24}$$

$$u_L(t) = u_L(t + L) \quad \text{for all } t \in (-\infty, \infty), \tag{3.25}$$

$$\|u(t) - u_L(t)\|_{X_2} \leqq M \exp(-\omega t) \quad \text{for all } t \geqq 0, \tag{3.26}$$

where M and ω are positive constants, and for each $0 < \delta \leqq \frac{1}{2}$, u_L is a Hölder continuous $\mathscr{D}(A^{1/2-\delta})$-valued function on $(-\infty, \infty)$.

Proof. The method of proof is similar to that of Theorem 4.2 of [10]. Let $u(x, t)$ be the global solution obtained in Theorem 3.1. Then, by setting $w = 0$ in (0.1), we have

$$\frac{1}{2}\frac{d}{dt}\|u\|_{X_2}^2 + a(u, u) \leqq (f, u) \quad \text{for almost all } t \in (0, \infty), \tag{3.27}$$

from which it follows that

$$\|u\|_{X_2}^2 \leqq \|u_0\|_{X_2}^2 \exp(-\omega_1 t) + C_{10} \operatorname*{ess\,sup}_{t \geqq 0} \|f(t)\|_{V'}^2, \tag{3.28}$$

where ω_1 and C_{10} are positive constants depending on Ω and μ. We next define

$$v_k(x, t) = u(x, t + kL) \quad \text{for } k = 0, 1, 2, \cdots. \tag{3.29}$$

Then, each v_k satisfies (0.1) on the interval $(-kL, \infty)$ and it is easily seen that

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|v_k - v_0\|_{X_2}^2 + C_{11}\|v_k - v_0\|_V^2 &\leqq |b(v_k - v_0, v_k - v_0, v_0)| \\ &\leqq C_{12}\|v_0\|_{L^3(\Omega)^3}\|v_k - v_0\|_V^2 \\ &\leqq C_{13}\|v_0\|_{X_2}^d \|v_0\|_{X_r}^{1-d}\|v_k - v_0\|_V^2 \\ &\qquad \text{for almost all } t \in (0, \infty), \end{aligned} \tag{3.30}$$

where C_{11}, C_{12}, and C_{13} are positive constants depending only on Ω and μ, and d is a number satisfying $d/2 + (1-d)/r = \frac{1}{3}$.

In the meantime, it follows from (3.20) that

$$\|v_0(t)\|_{X_r} = \|u(t)\|_{X_r} \leqq C_{14}(1 + E_0 + \Theta_2) \quad \text{for all } t \geqq 0, \tag{3.31}$$

where C_{14} is a positive constant independent of E_0 and Θ_2. By virtue of (3.28) and (3.31), we can choose a positive number $\tilde{C}$ in (3.23) so that (3.23) implies

$$C_{13}\|v_0(t)\|_{X_2}^d \|v_0(t)\|_{X_r}^{1-d} \leqq \tfrac{1}{2}C_{11} \quad \text{for all } t \geqq 0. \tag{3.32}$$

Hence, under conditions (3.1) and (3.23), we obtain from (3.30)

$$\frac{d}{dt}\|v_k - v_0\|_{X_2}^2 + C_{11}\|v_k - v_0\|_V^2 \leqq 0 \quad \text{for almost all } t \in (0, \infty), \tag{3.33}$$

which yields

$$\|v_k - v_0\|_{X_2} \leqq C_{15} \exp(-\omega_2 t) \quad \text{for all } t \geqq 0 \text{ and all } k \geqq 0, \tag{3.34}$$

where C_{15} and ω_2 are positive constants.

Now we find that

$$\begin{aligned} \|v_{k+m}(t) - v_m(t)\|_{X_2} &= \|v_k(t + mL) - v_0(t + mL)\|_{X_2} \\ &\leqq C_{15} \exp(-\omega_2 mL) \quad \text{for all } t \geqq 0 \text{ and all } k, m \geqq 0. \end{aligned} \tag{3.35}$$

Consequently, $\{v_k\}_{k=0}^{\infty}$ is a Cauchy sequence in $C([0, \infty); X_2)$. Let u_L be its limit. By the same argument as in [10], u_L satisfies (0.1) and (3.24)–(3.26). We omit the remaining details.

Acknowledgment. I thank Professors K. Hannsgen and R. Wheeler for their support. I am also grateful to the referee for constructive criticism.

REFERENCES

[1] L. CATTABRIGA, *Su un problema al contorno relativo al sistema di equazioni di Stokes*, Rend. Sem. Mat. Univ. Padova, 31 (1961), pp. 308-340.

[2] M. CRANDALL AND E. SOUGANIDIS, *Convergence of difference approximations of quasilinear evolution equations*, MRC Tech. Report 2711, University of Wisconsin, Madison, WI, 1984.

[3] G. DUVAUT AND J. L. LIONS, *Écoulement d'un fluide rigide viscoplastique incompressible*, C. R. Acad. Sci. Paris, 270 (1970), pp. 58-61.

[4] ———, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, New York, 1976.

[5] Y. GIGA, *Analyticity of the semigroup generated by the Stokes operator in L_r spaces*, Math. Z., 178 (1981), pp. 297-329.

[6] ———, *Domains of fractional powers of the Stokes operator in L_r spaces*, Arch. Rational Mech. Anal., 89 (1985), pp. 251-265.

[7] Y. GIGA AND T. MIYAKAWA, *Solutions in L_r of the Navier-Stokes initial value problem*, Arch. Rational Mech. Anal., 89 (1985), pp. 267-281.

[8] R. GLOWINSKI, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, Berlin, New York, 1984.

[9] J. KIM, *On the Cauchy problem associated with the motion of a Bingham fluid in the plane*, Trans. Amer. Math. Soc., 298 (1986), pp. 371-400.

[10] ———, *On the initial-boundary value problem for a Bingham fluid in a three dimensional domain*, Trans. Amer. Math. Soc., 304 (1987), pp. 751-770.

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

[12] J. NAUMANN AND M. WULST, *On evolution inequalities of Bingham type in three dimensions* II, J. Math. Anal. Appl., 70 (1979), pp. 309-325.

[13] W. PRAGER, *Introduction to Mechanics of Continua*, Ginn, Boston, New York, 1961.

[14] M. RENARDY, *Dense embedding of test functions in certain spaces*, Trans. Amer. Math. Soc., 298 (1986), pp. 241-243.

[15] R. TEMAM, *Navier-Stokes Equations*, North-Holland, Amsterdam, New York, 1984.

[16] W. L. WILKINSON, *Non-Newtonian Fluids*, Pergamon Press, New York, London, 1960.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 76-84, January 1990
© 1990 Society for Industrial and Applied Mathematics
005

GLOBAL BIFURCATION OF STEADY-STATE SOLUTIONS ON A BIOCHEMICAL SYSTEM*

CHUNQING LU†

Abstract. The differential equation $u''-(u/k\beta^2(u^2+2\varepsilon u+4k\varepsilon^2))=0$ with the boundary conditions $u(0)=u(1)=1$ governs the steady-state solutions from a mono-enzyme membrane model. It is proved that for a given $k>0$ there are at most three solutions for all $\varepsilon>0$ and for all $\beta>0$, and that there exists an $\varepsilon_*=\varepsilon_*(k)$, a value of ε, at which a pitchfork bifurcation occurs in the corresponding reaction-diffusion equations.

Key words. steady state, pitchfork, global bifurcation

AMS(MOS) subject classification. 92A09

1. Introduction. A model describing the diffusion and reaction of a substrate in a mono-enzymatic artificial membrane was established by Thomas (see Kernevez and Thomas [5]). The biochemical system is a membrane with the enzyme uricase linked to a support. The substrate is uric acid, and the cosubstrate is oxygen. The substrate and cosubstrate diffuse only within the membrane and they react in the presence of the enzyme (they are not parts of the membrane). The stoichiometric equation is:

$$\text{uric acid} + \text{oxygen} \xrightarrow{\text{uricase}} \text{allantoin} + \text{other products.}$$

Let $S=S(x, t)$ be the concentration of the substrate of the membrane. Then it satisfies the following reaction-diffusion equation:

$$S_t - D_S S_{xx} + R(S) = 0, \tag{1.1}$$

together with boundary conditions

$$S = S_0 \quad \text{at } x=0 \quad \text{and} \quad x = L \text{ (membrane thickness)},$$

and the given initial condition, where D_S is the coefficient of diffusion, a constant, and

$$R(S) = V_M S/[k_S + S(1+S/k_{SS})]$$

is the rate of the reaction where k_S is the Michaelis constant, k_{SS} the inhibition constant of S for the enzyme, and V_M is the maximal value of the reaction rate. Nondimensionalizing the equation, we obtain the following partial differential equation:

$$s_t - s_{xx} + \sigma F(s) = 0, \quad 0<x<1, \quad t>0 \tag{1.2}$$

with boundary conditions

$$s(0, t) = s(1, t) = s_0, \tag{1.2'}$$

and the given initial data $s(x, 0)$, where $s=S/k_S$ and $F(s)=s/(1+s+ks^2)$, and $\sigma=(V_M/k_S)(L^2/D_S)$, $k=k_S/k_{SS}$, and $s_0=S_0/k_S$ are positive constants. Then the steady-state equation associated with (1.2)-(1.2') is the two-point boundary value problem:

$$-s''+\sigma F(s)=0, \qquad 0<x<1, \tag{1.3}$$

$$s(0)=s(1)=s_0. \tag{1.3'}$$

* Received by the editors April 18, 1988; accepted for publication (in revised form) February 27, 1989.
† Institute of Software, Academia Sinica, P.O. Box 8718, Beijing, People's Republic of China.

Kernevez [4] showed that there exist at least three solutions for (1.3)-(1.3′) for large σ and for some s_0. In 1976, Brauner and Nicolaenko [1] studied the stability of the multiple steady-state solutions using the Crandall-Rabinowitz theorem for large σ based on the assumption that there are at most three solutions without a proof. In 1985, Lu [7] gave a rigorous proof: given $k>0$, there exists an $s_{0*}=s_{0*}(k)$ such that if $s_0>s_{0*}$, then there are at most three solutions for any $\sigma>0$. Lu also indicated that the method can also be applied to the case for large σ. In this paper, Lu continues his work of [6] and [7], and studies the global bifurcation for all s_0 and $\sigma>0$ for given k to prove that a pitchfork bifurcation point exists in the system.

Biologically, the results will explain that multiple stable steady states and a hysteresis phenomenon occur in a very simple biochemical system, such as this one, where diffusion and enzyme reaction interact, because s_0 is very large compared with the membrane thickness L. Mathematically, questions about numbers and stability of steady-state solutions of reaction-diffusion equations depend heavily upon their nonlinear terms, and there is no general way to handle them. Smoller and Wasserman [9] studied a case in which the nonlinearity is a polynomial. Hastings and McLeod [2] dealt with an exponential nonlinearity. In this paper, we consider a different nonlinearity from theirs—a class of rational functions.

2. Main results. We again apply the changes of variables used in [6] and [7]:

$$u(x)=s_0^{-1}s(x),\quad \beta=s_0\sigma^{-1/2},\quad \varepsilon=(2ks_0)^{-1}.$$

Then the given steady-state equation takes the form:

$$-u''+\frac{u}{k\beta^2(u^2+2\varepsilon u+4k\varepsilon^2)}=0,\qquad 0<x<1, \tag{2.1}$$

$$u(0)=u(1)=1, \tag{2.1′}$$

and the reaction-diffusion equation becomes

$$\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}+\frac{u}{k\beta^2(u^2+2\varepsilon u+4k\varepsilon^2)}=0,\quad 0<x<1,\quad t>0, \tag{2.2}$$

$$u(0,t)=u(1,t)=1,\qquad t>0, \tag{2.2′}$$

$$u(x,0)\text{ given.} \tag{2.2″}$$

The main results are the following three theorems.

THEOREM 1. *For any given $k>0$, there exist at most three solutions of* (2.1)-(2.1′) *for all positive β and ε.*

THEOREM 2. *For any given $k>0$, there exists an $\varepsilon_*=\varepsilon_*(k)$, a value of ε, such that for each $\varepsilon<\varepsilon_*$ there is a pair (β_1,β_2) depending on ε and k such that* (2.1)-(2.1′) *has exactly three solutions for $\beta\in(\beta_1,\beta_2)$, and only one solution for $\varepsilon\geqq\varepsilon_*$ and for any $\beta>0$.*

THEOREM 3. *Suppose that ε_* is chosen as in Theorem 2. If $\varepsilon>\varepsilon_*$, then the unique solution of* (2.1)-(2.1′) *is stable to* (2.2)-(2.2″); *if $\varepsilon\in(0,\varepsilon_*)$ and* (2.1)-(2.1′) *has three solutions, then two of them are stable and the other unstable.*

Remark 1. Let $u(x)$ be any solution of (2.1)-(2.1′) and $u(x,t)$ a solution of (2.2)-(2.2″). If for any given $\alpha>0$ there exists a $\delta>0$ such that $\|u(x,t)-u(x)\|<\alpha$ for all $t>0$ as long as $\|u(x,0)-u(x)\|<\delta$, then $u(x)$ is called stable to (2.2)-(2.2″); otherwise, it is unstable, where $\|\cdot\|$ is the C-norm.

Remark 2. It is observed from Theorems 2 and 3 that when the initial boundary value problem (2.2)-(2.2″) is treated as a local flow in a certain function space (cf. [10]), ε_* is a bifurcation point at which a pitchfork bifurcation occurs.

Remark 3. Because of the practical background of the equations, all parameters appearing in the u-equations are positive. Also, we assume that all solutions studied in the paper are positive as well. The existence of such solutions has been proved in [7].

3. Proofs of Theorems 1 and 2. We need some lemmas.

LEMMA 1. *If* $u(x)$ *solves* (2.1)-(2.1′), *then* $u(\frac{1}{2})=0$.

Proof. Multiply (2.1) by u', and then integrate both sides. This yields

$$\frac{[u'(x)]^2}{2}=\frac{1}{k\beta^2}\int_{u(1/2)}^{u(x)}\frac{u\,du}{u^2+2\varepsilon u+4k\varepsilon^2}+\frac{[u'(\frac{1}{2})]^2}{2}. \tag{3.1}$$

Then $u'(0)^2=u'(1)^2$. Since $u>0$, $u''>0$ and u is convex. Then $u'(0)=-u'(1)$. Let $x=\frac{1}{2}+z$ and $v(z)=u(\frac{1}{2}+z)$. Then $v(z)$ satisfies (2.1) in $(-\frac{1}{2},\frac{1}{2})$ with $v(-\frac{1}{2})=v(\frac{1}{2})=1$, and $v'(\frac{1}{2})=-v'(-\frac{1}{2})$. It is observed that $w(z)=v(-z)$ is also a solution of (2.1) with the same initial values as v. By the uniqueness of $v(z)$, $v(z)=v(-z)$, implying that $u'(\frac{1}{2})=0$.

Lemma 1 implies that the number of steady-state solutions will be determined by the number of values of $u(\frac{1}{2})$, which must satisfy certain conditions. To determine such conditions, as in [7], denote $u(\frac{1}{2})=1/y$, solve $u'(x)$ from (3.1), and integrate the obtained equation. Then introduce the change of variable: $\varepsilon+u=(\varepsilon+(1/y))e^{t^2}$. It follows that

$$2\left(\varepsilon+\frac{1}{y}\right)\int_0^v te^{t^2}\left[\ln\frac{(1+\varepsilon y)^2e^{2t^2}+(4k-1)\varepsilon^2y^2}{(1+\varepsilon y)^2+(4k-1)\varepsilon^2y^2}-2\varepsilon y\int_1^{(1+\varepsilon y)e^{t^2}-\varepsilon y}\frac{ds}{(s+\varepsilon y)^2+(4k-1)\varepsilon^2y^2}\right]^{-1/2}dt=\frac{1}{2\beta\sqrt{k}}, \tag{3.2}$$

where $v=v(\varepsilon,y)=(\ln(\varepsilon+1)-\ln(\varepsilon+1/y))^{1/2}$. We denote the function on the left-hand side of (3.2), the so-called response function, by $f(y;\varepsilon)$. In this paper it may be written as f or $f(y)$ depending on the context. Thus, the number of multiple steady states will be uniquely determined by the number of solutions of the function equation $f(y;\varepsilon)=(2\beta\sqrt{k})^{-1}$. Denote

$$G=G(y,t,s)=\ln\frac{(1+\varepsilon y)^2e^{2t^2}+(4k-1)\varepsilon^2y^2}{(1+\varepsilon y)^2+(4k-1)\varepsilon^2y^2}-2\varepsilon y\int_1^{(1+\varepsilon y)e^{t^2}-\varepsilon y}\frac{ds}{(s+\varepsilon y)^2+(4k-1)\varepsilon^2y^2}, \tag{3.3}$$

and $H=H(y,\varepsilon)=G(y,V(y,\varepsilon),\varepsilon)$. Then (3.2) becomes

$$f(y;\varepsilon)=2\left(\varepsilon+\frac{1}{y}\right)\int_0^v te^{t^2}G^{-1/2}\,dt. \tag{3.4}$$

Hence,

$$\frac{\partial f}{\partial y}=\frac{1+\varepsilon}{(1+\varepsilon y)yH^{1/2}}-\frac{2}{y^2}\int_0^v te^{t^2}G^{-1/2}\,dt-\left(\varepsilon+\frac{1}{y}\right)\int_0^v te^{t^2}G^{-3/2}\frac{\partial G}{\partial y}\,dt, \tag{3.5}$$

$$\begin{aligned}\frac{\partial^2 f}{\partial y^2}=&-\frac{2(1+\varepsilon)}{(1+\varepsilon y)y^2H^{1/2}}-\frac{1+\varepsilon}{2(1+\varepsilon y)yH^{3/2}}\left(\frac{\partial H}{\partial y}+\frac{\partial G}{\partial y}\bigg|_{t=v}\right)\\&+\frac{4}{y^3}\int_0^v te^{t^2}G^{-1/2}\,dt+\frac{3(1+\varepsilon y)}{2y}\int_0^v te^{t^2}G^{-5/2}\left(\frac{\partial G}{\partial y}\right)^2dt\\&+\int_0^v te^{t^2}G^{-3/2}\left(\frac{2}{y^2}\frac{\partial G}{\partial y}-\frac{1+\varepsilon y}{y}\frac{\partial^2 G}{\partial y^2}\right)dt.\end{aligned} \tag{3.6}$$

Since $\partial f/\partial y$ and $\partial^2 f/\partial y^2$ are too complicated, we introduce an auxiliary function $P = P(y;\varepsilon) = (1+\varepsilon y)^3 y((\partial^2 f/\partial y^2)+(2/y)(\partial f/\partial y))$.

For the remaining lemmas in this section, we present analytic proofs in detail only for $k=\frac{1}{4}$, because the proof for general k would be very tedious, as is pointed out in [6]. Also, the numerical results in [4] give good evidence that all properties of the response functions $f(y;\varepsilon)$ are demonstrated by the following lemmas. In the case $k=\frac{1}{4}$,

$$P = \frac{\varepsilon(y-2)-1}{yH^{3/2}} + 6\varepsilon^2 \int_0^v te^{t^2}(1-e^{-t^2})^2 G^{-5/2}\,dt. \tag{3.7}$$

LEMMA 2. *Function P has a unique zero in $(1,\infty)$ for any ε.*

Proof. It is observed from (3.7) that $P>0$ for $y \geqq (2+\varepsilon^{-1})$ and $\lim_{y\to 1^+} H(y:\varepsilon)=0$, and hence $\mathrm{Lim}_{y\to 1^+} P(y;\varepsilon) = -\infty$. This proves the existence of zeros of P in $(1,\infty)$. To see the uniqueness, we differentiate P with respect to y once:

$$\begin{aligned}\frac{\partial P}{\partial y} &= \frac{1+2\varepsilon}{y^2 H^{3/2}} + \frac{3\{\varepsilon^2(y-1)^2-(1+\varepsilon)[\varepsilon(y-2)-1]\}}{(1+\varepsilon y)^2 y^2(1+\varepsilon)H^{5/2}} \\ &\quad + \frac{30\varepsilon^3}{(1+\varepsilon y)^2}\int_0^v te^{t^2}(1-e^{-t^2})^3 G^{-7/2}\,dt.\end{aligned} \tag{3.8}$$

Then $\partial P/\partial y > 0$ for $y \in (1, 2+\varepsilon^{-1})$. The conclusion of the lemma follows readily.

We now proceed with the proof of Theorem 1. We mean that a function $Q(x)$ has a wiggle at a critical point $x=x_0$, or that x_0 is a wiggle-point of Q, if there exists a $\delta>0$ such that $Q'(x)<0$ ($Q'(x)>0$) for $x\in(x_0-\delta, x_0)$ and $Q'(x)>0$ ($q'(x)<0$) for $x\in(x_0, x_0+\delta)$. Therefore, to prove Theorem 1, we ought to show that $f(y;\varepsilon)$ is either monotonic or so-called "S-shaped," i.e., having two wiggles. We prove this by contradiction. Since $f(1;\varepsilon)=0$ and $f(y;\varepsilon)\to\infty (y\to\infty)$, f must be a function with an even number of wiggles. Suppose that f is neither monotonic nor does it have two wiggles. Then it must have at least four wiggles. Without loss of generality we assume that f has exactly four wiggles. Let the leftmost wiggle-point be x_1 at which f takes the local maximum and $f''\leqq 0$; the remaining wiggle-points in turn are $x_2<x_3<x_4$. By Lemma 2, given ε, there is at most one zero among $f''(x_i)$ $(i=1,2,3,4)$. We consider two subcases about $f''(x_i)$: (1) neither of them is zero; (2) only one is zero. If case (1) holds, then $f''(x_i)<0$ and $P(x_i)<0$ for $i=1,3$, and $f''(x_j)>0$ and $P(x_j)>0$ for $j=2,4$, which contradicts Lemma 2. If case (2) holds, say $f''(x_1)=0$, then $P(x_1)=0$. However, the fact that $f''(x_2)>0$ and $f''(x_3)<0$ implies $P(x_2)>0$ and $P(x_3)<0$. Again, it violates Lemma 2. Similarly, we can prove that it is impossible for any of the $f''(x_j)(j=2,3,4)$ to become zero. This means the previous assumption that f has at least four wiggles is wrong. The proof of Theorem 1 is complete.

The remaining lemmas are concerned with the proof of Theorem 2.

LEMMA 3. *There exists an ε_0, a value of ε, such that $f(y;\varepsilon)$ is a function with two wiggles for $\varepsilon\in(0,\varepsilon_0)$.*

LEMMA 4. *$f(y;0)$ is a function with exactly one wiggle, and takes its maximum at $x_0=2+\alpha$, where α is a constant.*

LEMMA 5. *Let α be chosen as in Lemma 4. Then $f'(y;\varepsilon)>0$ for all $y>1$ and for all $\varepsilon \geqq 1/\alpha$.*

Lemma 3 is the main result of [6].

Proof of Lemma 4. Elementary calculations show that

$$f(y;0) = \frac{\sqrt{2}}{y}\int_0^{\sqrt{\ln y}} e^{t^2}\,dt, \tag{3.9}$$

$$f'(y;0)=\frac{1}{y\sqrt{2\ln y}}-\frac{\sqrt{2}}{y^2}\int_0^{\sqrt{\ln y}} e^{t^2}\,dt, \tag{3.10}$$

$$f''(y;0)=\frac{-2}{y^2\sqrt{2\ln y}}-\frac{\sqrt{2}}{4y^2(2\ln y)^{3/2}}+\frac{2\sqrt{2}}{y^3}\int_0^{\sqrt{\ln y}} e^{t^2}\,dt. \tag{3.11}$$

Let $\Delta=\Delta(y)=f'(y;0)+2f'(y;0)/y$. Then $\Delta<0$ for all $y>1$, which means that $f''<0$ wherever $f'=0$; hence $f(y;0)$ is a function with only one wiggle at $x=x_0$, and takes the global maximum at x_0. Furthermore,

$$\begin{aligned} 2\sqrt{2\ln 2}\,f'(2;0) &= 1-\sqrt{\ln 2}\int_0^{\sqrt{\ln 2}} e^{t^2}\,dt \\ &= 1-\sqrt{\ln 2}\int_0^{\sqrt{\ln 2}} \sum_0^{\infty}\frac{t^{2n}}{n!}\,dt \\ &> 1-\left\{\ln 2+\frac{(\ln 2)^2}{3}+\frac{(\ln 2)^3}{10}\sum_0^{\infty}\left[\frac{\ln 2}{4}\right]^n\right\}. \end{aligned} \tag{3.12}$$

Using $\ln 2<0.7$, we obtain $2\sqrt{2\ln 2}\,f'(2;0)>0.09$, and hence $f(2;0)>0$. Similarly, we prove $f'(e;0)<0$. Therefore, the wiggle-point of $f(y;0)$, $x_0=2+\alpha$ for some $\alpha\in(0,e-2)$. This proves the lemma.

Proof of Lemma 5. Differentiating $f'(y;\varepsilon)$ with respect to ε yields

$$\frac{\partial^2 f}{\partial\varepsilon\partial y}=\frac{2(y-1)}{(1+\varepsilon y)^3 yH^{3/2}}+\frac{6\varepsilon}{(1+\varepsilon y)^3}\int_0^{v} te^{t^2}(1-e^{-t^2})^2G^{-5/2}\,dt. \tag{3.13}$$

It is seen from (3.13) that given $y>1$, $f'(y;\varepsilon)$ increases as ε does. Thus $f'(y;\varepsilon)>0$ for $y\in(1,2+\alpha]$ and for all $\varepsilon>0$, because $f'(y;0)>0$ for $y\in(1,2+\alpha)$ by the preceding lemma. If $y>2+\alpha$, then $y>2+\varepsilon^{-1}$, and hence $P(y;\varepsilon)>0$. Therefore, $f'(y;\varepsilon)>0$ for $y>2+\alpha$, for otherwise $f'\leqq 0$ leads $f''>0$; i.e., f would reach its minimum first on the left, which is impossible. This proves Lemma 5.

The geometrical meaning of Theorem 2 is as follows. There exists a value of ε, $\varepsilon_*>0$ such that $f(y;\varepsilon)$ has exactly two wiggles for $\varepsilon\in(0,\varepsilon_*)$, and that $f'(y;\varepsilon)\geqq 0$ for $\varepsilon\geqq\varepsilon_*$. To prove this, we define two subsets M and N on the real line as follows:

$$M=\{\varepsilon\,|\,f(y;\varepsilon)\text{ with only two wiggles for all }y>1\}, \tag{3.14}$$

$$N=\{\varepsilon\,|\,\varepsilon>0\text{ and }f'(y;\varepsilon)<0\text{ for some }y>1\}. \tag{3.14'}$$

LEMMA 6. *M is an open interval, and $M=N$.*

Proof. We first prove $M=N$. It is trivial that $M\subset N$ by the definition about the wiggle-point. Take an $\varepsilon\in N$, so that $f'(y';\varepsilon)<0$ for some $y'>1$. Since $f'(y;\varepsilon)\to\infty$ as $y\to 1+$ and $f(y;\varepsilon)\to\infty$ as $y\to\infty$, therefore $f'(y^-;\varepsilon)>0$ and $f'(y^+;\varepsilon)<0$ for some y^- and y^+, where $1<y^-<y'<y^+$. This leads us to the fact that f has at least two wiggles. Then, from the proof of Theorem 1, such $\varepsilon\in M$; hence $N\subset M$. Obviously, N is open and so is M. Also, M is nonempty and bounded from Lemmas 3 and 5. Note that the complement of N, $N^{\subset}=\{\varepsilon\,|\,f'(y;\varepsilon)\geqq 0\text{ for all }y>1\}$ is a continuum because $\varepsilon'\in N^{\subset}$ implies that $(\varepsilon',\infty)\subset N^{\subset}$ by $\partial^2 f/\partial\varepsilon\partial y>0$ from (3.13). Therefore, M is an open interval containing $(0,\varepsilon_0)$, where ε_0 is given by Lemma 3. This proves the lemma.

Now set $\varepsilon_*=\operatorname{Sup} M$. We see from the proof of Lemma 6 that $\varepsilon_*\in N^{\subset}$ and $f'(y;\varepsilon)\geqq 0$ for all $\varepsilon\geqq\varepsilon_*$, and that $\varepsilon\in M$ and f has exactly two wiggles for $\varepsilon<\varepsilon_*$. In fact, the next lemma shows that ε_* bifurcates the numbers of the steady-state solutions.

LEMMA 7. *$f'(y;\varepsilon)>0$ for all $\varepsilon>\varepsilon_*$, and $f'(y;\varepsilon)$ has a unique zero y_** (cf. Fig. 1).

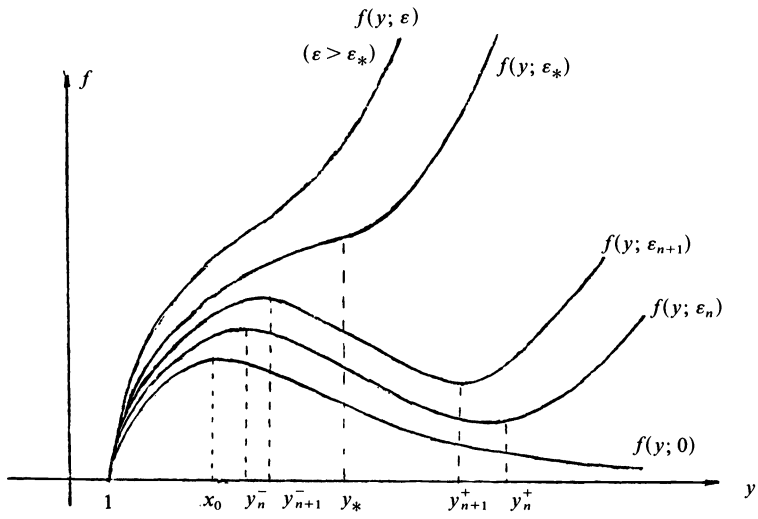


FIG. 1

Proof. Take a monotonically increasing sequence $\{s_n\}$ $n = 1, 2 \cdots$ such that $\varepsilon_n \uparrow \varepsilon_*$ as $n \to \infty$. Let $y_n^- < y_n^+$ be the two wiggle-points of $f(y;\varepsilon_n)$. Then $f'(y_n^-;\varepsilon_n) = f'(y_n^+;\varepsilon_n) = 0$, $f''(y_n^-;\varepsilon_n) \leqq 0$ and $f''(\varepsilon_n^+;\varepsilon_n) \leqq 0$ for all n. Also, $f'(y_{n+1}^-;\varepsilon_n) < 0$ and $f'(y_{n+1}^+;\varepsilon_n) < 0$ by $\partial^2 f/\partial\varepsilon\,\partial y > 0$. Thus $[y_n^-, y_n^+] \supset [y_{n+1}^-, y_{n+1}^+]$; hence there exist $y_*^- \leqq y_*^+$ such that $y_n^- \uparrow y_*^-$ and $y_n^+ \uparrow y_*^+$ as $n \to \infty$. Since f' and f'' are continuous on any subsets of the $y-\varepsilon$ plane, $[y', y''] \times [\varepsilon', \varepsilon'']$ where y', y'', ε', and ε'' are arbitrary real numbers, $f'(y_n^-;\varepsilon_n) \to f'(y_*^-;\varepsilon_*)$ and $f'(y_n^+;\varepsilon_n) \to f'(y_*^-;\varepsilon_*)$ as $n \to \infty$. Therefore $f'(y;\varepsilon_*)$ has zeros y_*^- and y_*^+. Next we prove the uniqueness of the zero by contradiction. Suppose that y_1 is the leftmost zero of $f'(y;\varepsilon_*)$, and y_2 the nearest zero to y_1. Then $f''(y_1;\varepsilon_*) \leqq 0$. If $f''(y_1;\varepsilon_*) < 0$, then $f'(y_1;\varepsilon_*) = 0$ implies $f' < 0$ for some $y > y_1$; hence $\varepsilon_* \in N$. This violates $\varepsilon_* \notin M$. If $f''(y_1;\varepsilon_*) = 0$, then $f''(y_2;\varepsilon_*) > 0$ by Lemma 2, and hence $f' < 0$ for some $y < y_2$. Again, it violates $s_* \notin M$. The lemma is proved.

Since $\partial^2 f/\partial\varepsilon\,\partial y > 0$, $f'(y;\varepsilon) > 0$ for all $\varepsilon > \varepsilon_*$ and for all $y > 1$. For $\varepsilon \in (0, \varepsilon_*)$, let $y_-(\varepsilon) < y_+(\varepsilon)$ be the two wiggle-points. We set $\beta_1 = [2\sqrt{k} f(y_-;\varepsilon)]^{-1}$ and $\beta_2(\varepsilon) = [2\sqrt{k} f(y_+;\varepsilon)]^{-1}$. Then the conclusions of Theorem 2 follow immediately.

4. Proof of Theorem 3. We will use the Morse Index Theorem to investigate the eigenvalue problem. For convenience, let $Q = Q(u) = \sigma F(u)$, where $F(u)$ is given in § 1. In this paper we use the following definition.

DEFINITION. A number λ and a nontrivial function $v(x)$ are called an eigenvalue and an eigenfunction (corresponding to λ) associated with $u(x)$, a solution of (2.1)-(2.1'), respectively, if they satisfy the following boundary value problem:

$$v'' - Q_u v = \lambda v, \qquad 0 < x < 1, \tag{4.1}$$

$$v(0) = v(1) = 0. \tag{4.1'}$$

It is well known that there exists a unique positive eigenfunction on $[0, 1]$, and that the corresponding eigenvalue is real and simple (cf. [8]). They are called fundamental or principal eigenfunction and eigenvalue. In this paper we always mean such eigenfunction and eigenvalue. Also, it is known that the parabolic partial differential equation (2.2)-(2.2') may be treated as a local flow in a certain function space W, and that the steady-state solution $u(x)$ is stable for $\lambda < 0$, and unstable for $\lambda > 0$ in the sense of Lyapunov.

Given $u(x)$, a solution of (2.1)-(2.1'), we introduce an operator $L=(d^2/dx^2)-Q_u$ on W. Define $L[v]=v''-Q_uv$ for $v\in W$. Then the Morse index of L on the interval $(0, 1]$ equals the number of positive eigenvalues associated with $u(x)$. Therefore, $u(x)$ is stable when the index is zero; $u(x)$ is unstable when the index is 1.

Remark. Suppose that $w(x)\in W$ and satisfies

$$L[w]=0, \qquad 0<x<1, \tag{4.2}$$

$$w(0)=0, \qquad w'(0)\neq 0. \tag{4.2'}$$

Then the number of zeros of functions of this kind on the interval $(0, 1]$ does not depend on choices of $w(x)$. We now introduce the Morse Index Theorem in this simple case.

MORSE INDEX THEOREM. *The Morse Index of the operator L is finite and equal to the number of zeros of any function $w(x)$ satisfying* (4.2)-(4.2') (*cf.* [3]).

To apply the Morse Index Theorem we construct the following functions:

$$w(x)=\begin{cases} u'(x)\displaystyle\int_0^x (u')^{-2}\,dt, & 0\leqq x<\tfrac{1}{2}, \\ u'(x)\left[A+\displaystyle\int_1^x (u')^{-2}\,dt\right], & \tfrac{1}{2}\leqq x<1, \end{cases} \tag{4.3}$$

where

$$A=\frac{2}{u''(\frac{1}{2})}\left[\frac{1}{u'(0)}+\int_0^{1/2}\frac{u'(\frac{1}{2})-u''(t)}{[u'(t)]^2}\,dt\right]. \tag{4.4}$$

It can be proved that $w(x)$ given by (4.3) is a C^2 function and satisfies (4.2)-(4.2'). All we need is to compute their zeros for different cases of $u(x)$.

LEMMA 8. *The number of zeros on* $(0, 1]$ *of* $w(x)$ *given by* (4.3) *is at most one. It is zero for* $A<0$, *and one for* $A>0$.

LEMMA 9. *Suppose that* $u(x)$ *solves* (2.1)-(2.1') *and* $u(\frac{1}{2})=1/y$. *If* $f'(y;\varepsilon)\neq 0$, *then* $u(x)$ *is nondegenerate, i.e., the eigenvalue of the operator L is nonzero. Furthermore, if* $f'(y;\varepsilon)>0$, *then* $u(x)$ *is stable; if* $f'(y;\varepsilon)<0$, *then* $u(x)$ *is unstable.*

Proof of Lemma 8. Since the solutions of (2.1)-(2.1') we studied are positive, $u''(x)>0$; hence $u'(x)<0$ in $[0, \frac{1}{2})$ and $u'(x)>0$ in $(\frac{1}{2}, 1]$. We observe from (4.3) that if $A<0$, then $w(x)<0$ for all $x\in(0, 1]$, and that if $A>0$, then $w(1)=Au'(1)>0$ implies that $w(x)$ has at least one zero on $(\frac{1}{2}, 1]$. On the other hand, since the function $\{A+\int_1^x [u'(t)]^{-2}\,dt\}$ is monotonically increasing and $u'>0$ for $x\in(\frac{1}{2}, 1]$, $w(x)$ has at most one zero on $(\frac{1}{2}, 1]$. This proves the lemma.

Proof of Lemma 9. For simplicity, let $M(u)=\int Q(u)\,du$. Then the response function $f(y;\varepsilon)$ becomes

$$f(y;\varepsilon)=\frac{1}{\beta}\int_{1/y}^1 \frac{du}{\sqrt{M(u)-M(1/y)}}. \tag{4.5}$$

Denote $u(\frac{1}{2})=\eta$ $(=1/y)$ and $T(\eta)=\beta f(1/\eta;\varepsilon)$. Then

$$T(\eta)=\int_0^{1-\eta}\frac{ds}{\sqrt{M(s+\eta)-M(\eta)}}, \tag{4.6}$$

$$T'(\eta) = -\frac{1}{\sqrt{M(1)-M(\eta)}} - \int_\eta^1 \frac{M'(u)-M'(\eta)}{2[\sqrt{M(u)-M(\eta)}]^3}\,du$$

$$= \sqrt{2}\left[-\frac{1}{u'(1)} - \int_{1/2}^1 \frac{u''(t)-u''(\tfrac{1}{2})}{[u'(t)]^2}\,dt\right] \tag{4.7}$$

$$= \sqrt{2}\left[\frac{1}{u'(0)} + \int_0^{1/2} \frac{u''(\tfrac{1}{2})-u''(t)}{[u'(t)]^2}\,dt\right],$$

because $u'' = Q(u)$ and $u'^2 = 2[M(u)-M(\eta)]$. It turns out from (4.4) that

$$A = \frac{\sqrt{2}\,T'(\eta)}{u''(\tfrac{1}{2})} = -\frac{\sqrt{2}\,\beta f'(y;\varepsilon)}{u''(\tfrac{1}{2})y^2}. \tag{4.8}$$

We have checked that $L[w]=0$ and $L[u']=0$, and that u' and w are linearly independent on $(0,1)$. Therefore, the general solution of the second-order differential equation $L[z]=0$ can be expressed by $z = c_1u' + c_2w$, where c_1 and c_2 are arbitrary constants. Suppose that $Y(0)=Y(1)=0$ and $L[Y]=0$. Then $Y\equiv 0$ on $[0,1]$ by the uniqueness of the solution of $L[z]=0$, which implies that the eigenvalue of L is nonzero. If $f'>0$, then $A<0$ by (4.8), and the Morse Index is zero by Lemma 8. Hence, by the Morse Index Theorem, the eigenvalue of L is negative. Therefore, $u(x)$ is stable. Similarly, if $f'<0$, then $u(x)$ is unstable.

The proof of Theorem 2 has already shown that (i) if $\varepsilon > \varepsilon_*$ then $f'(y;\varepsilon)>0$ for all $y \geqq 1$ and hence the unique steady-state solution is stable; (ii) if $0<\varepsilon<\varepsilon_*$ and $\beta \in (\beta_1, \beta_2)$, then there are three solutions u_1, u_2, and u_3 of (2.1)-(2.1′). Let $u_i(\frac{1}{2}) = 1/y_i$ $(i=1,2,3)$; hence $f'(y_j;\varepsilon)>0$ for $j=1,3$ and $f'(y_2;\varepsilon)<0$. By Lemma 9, we prove Theorem 3 immediately.

We can apply the Conley Index Theory to get more information about the global structure of the multiple steady states. It is seen that the Conley Index of u_2, $h(u_2)=\Sigma^1$, is a pointed circle, and the Conley Index of u_j, $h(u_j)=\Sigma^0$, is a pointed zero sphere $(j=1,3)$. Then there exist solutions v_1 and v_2 of (2.2)-(2.2′) connecting u_2 to u_1 and u_2 to u_3, respectively [10, Thm. 22.33], namely, $v_1(x,t)\to u_1(x)$ and $v_2(x,t)\to u_3(x)$ as $t\to\infty$, $v_1(x,t)\to u_2(x)$ and $v_2(x,t)\to u_2(x)$ as $t\to-\infty$ uniformly on $[0,1]$. Suppose that $u_*(x,t)$ solves (2.2)-(2.2′) with $u_*(x,0)=u_*(x)$. We claim that if $u_2(x)<u_*(x)<u_1(x)$ for $x\in(0,1)$, then u_* lies in the stable manifold of u_1, i.e., $u_*(x,t)\to u_1(x)$ as $t\to\infty$, while if $u_3(x)<u_*(x)<u_2(x)$ for $x\in(0,1)$, then u_* lies in the stable manifold of u_3. To see this, suppose, for example, that $u_1(x)>u_*(x)>u_2(x)$ for $x\in(0,1)$. Then $u_1(x)>u_*(x,t_0)>u_2(x)$ for some $t_0>0$ and $x\in(0,1)$, and it follows that $\partial u_*(0,t_0)/\partial x<0$ (because $u_*(x,t_0)<u_1<1$ and $u_*(0,t_0)=1$), and $\partial u_*(1,t_0)/\partial x>0$. Thus if v_1 is the solution connecting u_2 to u_1, then by the fact that $v_1(x,t)\to u_2$ as $t\to-\infty$ uniformly on $[0,1]$ we see that $v_1(x,t_0-t_1)<u_*(x,t_0)$ for some $t_1>0$. Let $w(x,t)\equiv v_1(x,t-t_1)$. Then $w(0,t)=w(1,t)=1$ and $w(x,t_0)<u_*(x,t_0)$. By the comparison theorem we obtain $w(x,t)<u_*(x,t)$ for all $t\geqq t_0$ and for $x\in(0,1)$. Meanwhile, $w(x,t)\to u_1(x)$ uniformly as $t\to\infty$, so that the same is true for $u_*(x,t)$; in other words, u_* lies in the stable manifold of u_1. In the same way, we can prove that if $u_3<u_*<u_2$, then u_* lies in the stable manifold of u_3.

We can also investigate the stability using the maximum principle to prove that the region of attraction of u_1 includes all initial distributions $v(x,0)$ satisfying $u_2(x)<v(x,0)\leqq 1$, while that of u_3 includes all $v(x,0)$ satisfying $\underline{u}(x)\leqq v(x,0)<u_2(x)$, where $\underline{u}(x) = (\cosh(x-\frac{1}{2})/2k\beta\varepsilon)/(\cosh(1/4k\beta\varepsilon)$ is a lower solution of (2.1) (cf. [7]). This is illustrated in Fig. 2.

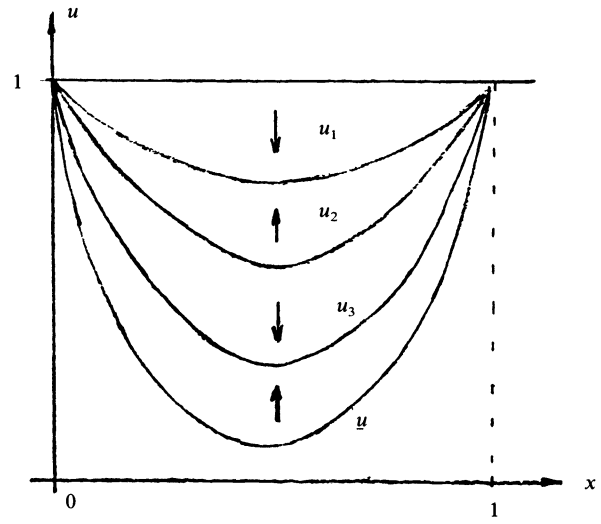


FIG. 2

REFERENCES

[1] C. M. BRAUNER AND B. NICOLAENKO, *Singular Perturbation, Multiple Solutions and Hysteresis in a Nonlinear Problem*, Lecture Notes in Math. 594, Springer-Verlag, Berlin, New York, 1977, pp. 50–76.

[2] S. P. HASTINGS AND J. B. MCLEOD, *The number of solutions of an equation from catalysis*, MRC Tech. Summary Report 2597, Mathematics Research Center, University of Wisconsin, Madison, WI, 1983.

[3] P. HERMANN, *Differential Geometry and Calculus of Variations*, Math-Sci Press, Brookline, MA, 1977.

[4] J. P. KERNEVEZ, *Enzyme Mathematics*, North-Holland, Amsterdam, New York, 1980.

[5] J. P. KERNEVEZ AND D. THOMAS, *Numerical analysis and control of some biochemical systems*, Appl. Math. Optim., 1 (1975), pp. 222–258.

[6] C. LU, *Multiple steady states in a biochemical system*, SIAM J. Math. Anal., 18 (1987), pp. 1771–1783.

[7] ———, *Multiple steady states and their stability in a biochemical system*, Ph.D. dissertation, State University of New York, Buffalo, NY, February 1986.

[8] D. H. SATTINGER, *Topics in Stability and Bifurcation Theory*, Lecture Notes in Mathematics 309, Springer-Verlag, Berlin, New York, 1973.

[9] J. SMOLLER AND A. WASSERMAN, *Global bifurcation of steady-state solutions*, J. Differential Equations, 39 (1981), pp. 269–290.

[10] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, New York, 1983.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 85-122, January 1990
© 1990 Society for Industrial and Applied Mathematics
006

SINGULAR LIMIT ANALYSIS OF STABILITY OF TRAVELING WAVE SOLUTIONS IN BISTABLE REACTION-DIFFUSION SYSTEMS*

Y. NISHIURA†, M. MIMURA†, H. IKEDA‡, AND H. FUJII§

Abstract. The stability properties of the traveling front solutions to bistable reaction-diffusion systems in which there are big differences in both the diffusion rates and the reaction rates between two species are studied. In contrast to the scalar case, this bistable system has multiple existence of traveling waves in the appropriate region of parameters. Each wave can be constructed by using a singular perturbation method, and its stability or instability is determined by a simple algebraic quantity appearing in its construction: namely, the sign of the Jacobian of inner and outer matching conditions. The singular limit approach (which is quite different from formal limiting arguments) adopted in this paper is rigorous and very useful in the study of stability problems of singularly perturbed solutions.

Key words. stability, traveling wave, singular perturbation, reaction-diffusion system

AMS(MOS) subject classifications. 35B25, 35B40, 35K57

1. Introduction.

Bistable media are one of the basic machineries that create a variety of propagating patterns. Especially, traveling front waves describing the transition from one stable state to the other are the most essential and interesting ones for such media. For a two-component model system, we meet the following equations:

$$(P)_{\varepsilon,\tau} \qquad \begin{aligned} \varepsilon\tau u_t &= \varepsilon^2 u_{zz} + f(u, v), \\ v_t &= v_{zz} + g(u, v), \end{aligned} \qquad (t, z) \in (0, \infty) \times \mathbf{R}.$$

Here ε and τ are real parameters, where $\varepsilon\tau$ and ε/τ are, respectively, the ratios of the rates of reaction and diffusion of the quantities u and v. Suppose that ε is sufficiently small. When τ is of the order ε, the diffusion rates of u and v are of the same order, but u reacts much faster than v. On the other hand, when τ is of order $1/\varepsilon$, u reacts with the same order as v, although there is a big difference between the diffusion rates of u and v. The qualitative information about f and g is depicted in Fig. 1: f is a cubic-like function, and $g = 0$ intersects with $f = 0$ at three points $E_\pm$ and E_0.

Note that E_+ and E_- are stable constant solutions of $(P)_{\varepsilon,\tau}$. It is natural to expect the existence of traveling fronts connecting E_- to E_+. For such an existence problem, we know at least numerically that, when $\tau = O(\varepsilon)$, there occurs a multiple existence of traveling front solutions. On the other hand, when $\tau = O(1/\varepsilon)$, there is one solution [9] that is proved to be stable in [11]. We can imagine from this that a certain transition process might happen to the structure of solutions when τ varies between two extreme values. In fact, Ikeda, Mimura, and Nishiura [10] have recently studied the case $\tau = O(1)$, where there are differences in *both* reaction and diffusion rates between u and v, namely, u reacts much faster than v but diffuses much slower than v. It is noteworthy that the number of traveling fronts connecting E_- to E_+ depends crucially on the parameter τ. More precisely, on the one hand, there exists a unique traveling front solution for large τ and, on the other hand, there exist at least three solutions for small τ. In fact, when f and g are specified, respectively, as

$$f(u, v) = u(1-u)(u-a) - v \quad \text{and} \quad g(u, v) = u - \gamma v \tag{1.1}$$

with constants a and γ, Fig. 2 shows a typical situation of the dependency of traveling

* Received by the editors June 27, 1988; accepted for publication (in revised form) April 3, 1989.
† Department of Mathematics, Hiroshima University, Hiroshima 730, Japan.
‡ Department of Mathematics, Toyama University, Toyama 930, Japan.
§ Institute of Computer Sciences, Kyoto Sangyo University, Kyoto 603, Japan.

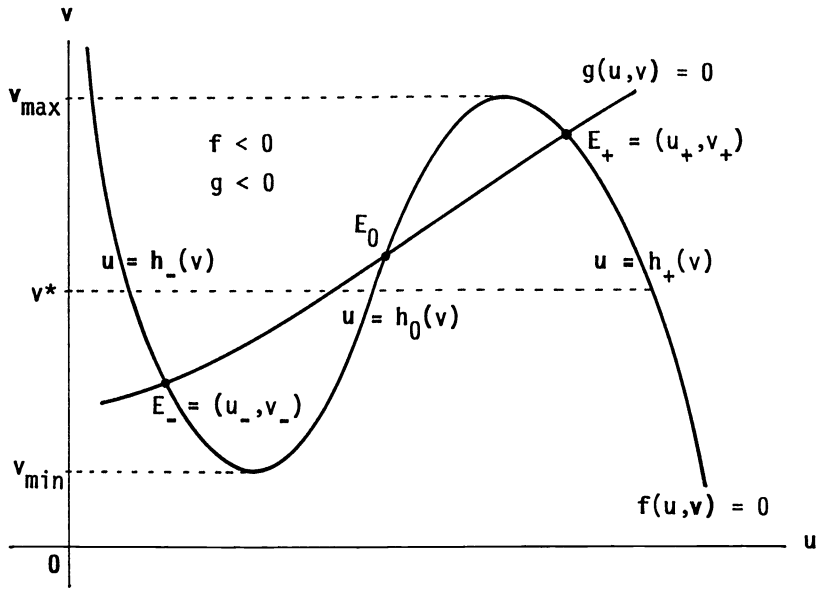


FIG. 1. *Functional forms of* $f = 0$ *and* $g = 0$.

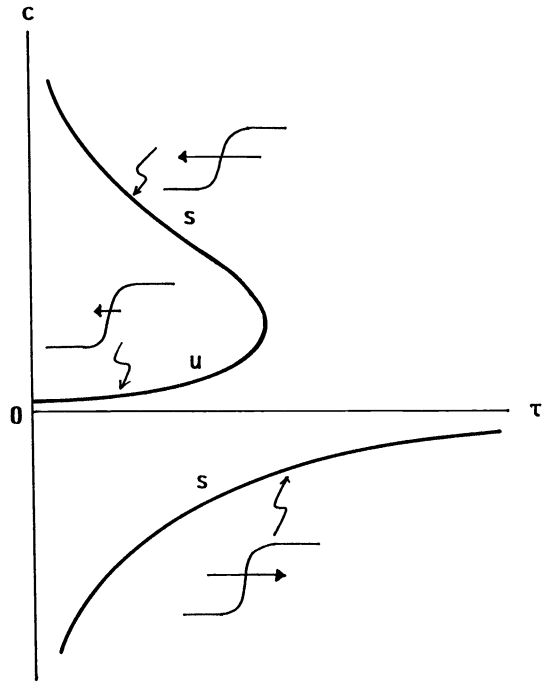


FIG. 2. *Global bifurcation diagram of traveling front solutions in* (τ, c)*-plane, where* s *(respectively,* u*) represents the stable (respectively, unstable) branch.*

front solutions on τ. This contrasts with the scalar reaction-diffusion equation of bistable type (see, e.g., Fife and McLeod [6]) where the front solution is always *unique* and *stable.* Multiple existence of traveling front solutions is also shown by Rinzel and Terman [16] for the bistable FitzHugh-Nagumo system with piecewise nonlinearity.

The aim of this paper is twofold concerning the stability properties of traveling front solutions of $(P)_{\varepsilon,\tau}$ when ε is sufficiently small and $\tau = O(1)$. First, we clarify the stability properties of the front solutions by using the Singular Limit Eigenvalue Problem (SLEP) method originated in [13] (see Theorems 3.1 and 3.2). Second, in Theorem 4.1 we give an alternative form of the stability criterion of Theorems 3.1 and 3.2, which essentially stems from the *geometrical* nature of the construction of singular limit traveling front solutions in § 2. Loosely speaking, our results can be summarized as follows.

MAIN THEOREM. *The linearized eigenvalue problem at a traveling front solution has a unique real simple eigenvalue besides the translation free zero eigenvalue and the rest of the spectrum lies strictly inside of the left halfplane independently of the parameters ε and τ. This critical eigenvalue is obtained as a zero of a scalar equation called the* SLEP *equation (see* (3.68) *and* (3.69)), *and its sign determines the stability of the traveling front, namely, positive (respectively, negative) means unstable (respectively, asymptotically stable in the orbital sense). Moreover, the sign of the critical eigenvalue is equal to that of the Jacobian of the C^1-matching conditions for outer and inner solutions of a singularly perturbed traveling front wave (see* (2.16) *and* (4.25)).

The last statement is close to the spirit of Evans' works [3], [4], which relate the stability of nerve pulse to the intersecting manner of stable and unstable manifolds. This geometrical interpretation of the stability criterion is very useful in a practical sense. Namely, when we construct a lowest-order approximation to a traveling front solution as $\varepsilon \downarrow 0$, we can judge its stability *simultaneously.* In fact, by looking at the schematic diagram Fig. 2, we can say from the construction of each solution that, except for the limit point, the upper and the lower branch are stable, while the middle one is unstable. This is quite reasonable from a bifurcation point of view. The details will be shown in § 4.

The idea of the SLEP method is to derive a limiting linearized eigenvalue problem as $\varepsilon \downarrow 0$ without losing information coming from the transition layer. It turns out that a Dirac point mass (in the one-dimensional case) appears in the limit of $\varepsilon \downarrow 0$ after an appropriate scaling, and the coefficient of it is determined by the global geometrical quantities of f and g. After some computation, the whole problem is reduced to solving a transcendental equation, or more geometrically, to finding intersection points of a straight line and a convex curve, which tells us the limiting location of dangerous eigenvalues to the stability (see (3.69)). See [12]-[15] for the details of the SLEP method and its applications.

What we would like to emphasize in this paper is that the singular limit analysis as $\varepsilon \downarrow 0$ (which is essentially different from $\varepsilon = 0$) sheds light on the world of $\varepsilon > 0$.

Finally, Jones and Gardner [17] have notified us that a topological approach may work to solve the same stability problem as above.

We impose the following assumptions on the nonlinearities of f and g (see Fig. 1).

(A0) f and g are smooth functions of u and v on some open set Q in $\mathbf{R}^2$.

(A1) $f = 0$ is S-shaped and consists of three branches $u = h_-(v)$, $h_0(v)$, and $h_+(v)$ $(h_-(v) \leqq h_0(v) \leqq h_+(v))$, while $g = 0$ intersects once with each branch at $E_- = (u_-, v_-)$, E_0, and $E_+ = (u_+, v_+)$ $(v_- < v_+)$, respectively, as in Fig. 1. The signs of f and g are both negative in the upper region of the curves $f = 0$ and $g = 0$.

(A2) $J(v) \equiv \int_{h_-(v)}^{h_+(v)} f(u, v)\, du$ has a unique isolated zero at $v^* \in (v_{\min}, v_{\max})$.

(A3) $f_u(h_\pm(v), v) < 0$ for $v \in [v_-, v_+]$,

$g(h_-(v), v) < 0 < g(h_+(v), v)$ for $v \in (v_-, v_+)$,

$g_v(h_\pm(v), v) < 0$ for $v = v_\pm$,

$$\frac{\partial(f, g)}{\partial(u, v)}(h_\pm(v), v) > 0 \quad \text{for } v \in [v_-, v_+].$$

(A4) $f_v(u, v) < 0$ for $(u, v) \in \{(u, v) \mid h_-(v) \leqq u \leqq h_+(v),\ v_- \leqq v \leqq v_+\}$,

$g_u(u, v) > 0$ at $(u, v) = (u_\pm, v_\pm)$.

Remark 1.1. The assumption for the sign of $(\partial(f, g)/\partial(u, v))(h_\pm(v), v)$ is equivalent to

$$\frac{d}{dv} g(h_\pm(v), v) < 0 \quad \text{for } v \in [v_-, v_+],$$

since, from $f(h_\pm(v), v) = 0$,

$$\frac{d}{dv} g(h_\pm(v), v) = \left. \frac{f_u g_v - f_v g_u}{f_u} \right|_{(u,v)=(h_\pm(v),v)}$$

holds.

Throughout this paper, we will use the following function spaces and notation. Let $I = \mathbf{R}_-$, $\mathbf{R}_+$ or $\mathbf{R}$, ρ, and σ be positive numbers, and let n be a nonnegative integer;

$$X^n_{\rho,\sigma}(I) \equiv \left\{ u \in C^n(I) \;\middle|\; \|u\|_{X^n_{\rho,\sigma}(I)} \equiv \sum_{i=0}^{n} \sup_{x \in I} \left| e^{\rho|x|} \left(\sigma \frac{d}{dx} \right)^i u(x) \right| < \infty \right\},$$

$$C_{\text{unif}}(I) \equiv \{u \mid u \text{ is bounded and uniformly continuous on } I\},$$

$$\mathbf{C}_{\text{unif}}(I) \equiv C_{\text{unif}}(I) \times C_{\text{unif}}(I),$$

$$H^n_\rho(I) \equiv \left\{ u \in H^n(I) \;\middle|\; \|u\|_{H^n_\rho(I)} \equiv \left\{ \sum_{i=0}^{n} \int_I \left| e^{\rho|x|} \left(\frac{d}{dx} \right)^i u(x) \right|^2 dx \right\}^{1/2} < \infty \right\}$$

where $H^n(I)$ is the usual Sobolev space on I;

$$L^2_\rho(I) \equiv H^0_\rho(I),$$

$$(H^n_\rho)^*(I) \equiv \text{the dual space of } H^n_\rho(I),$$

$\langle \cdot, \cdot \rangle$; L^2 inner product. $\langle \cdot, \cdot \rangle_x$ is also used to specify the independent variable;

$$\mathbf{C}_\mu \equiv \{\lambda \in \mathbf{C} \mid \operatorname{Re} \lambda > -\mu,\ \mu \in \mathbf{R}_+\},$$

$\sigma_e(\mathscr{L})$ identically equals the set of the essential spectrum of the operator $\mathscr{L}$; $C^n_{\text{c.u.}}(I)$ identically equals the uniform convergence on any compact subset of I in $C^n(I)$-sense.

2. Construction of traveling front solutions. In this section we will summarize the existence results of traveling front solutions studied in the previous paper [10]. Introducing the traveling coordinate $x = z + ct$, we find that traveling front solutions with velocity c satisfy

$$\begin{aligned} &\varepsilon^2 u_{xx} - \varepsilon c \tau u_x + f(u, v) = 0, \\ &v_{xx} - c v_x + g(u, v) = 0, \end{aligned} \qquad x \in \mathbf{R} \tag{2.1}$$

with boundary conditions

$$u(\pm\infty) = u_\pm, \qquad v(\pm\infty) = v_\pm. \tag{2.2}$$

To avoid the phase ambiguity, we impose the following condition on $u(x)$:

$$u(0) = \alpha \tag{2.3}$$

where α is an arbitrarily fixed value in some interval (see § 2.2). Moreover, we put

$$v(0) = \beta \tag{2.4}$$

for $\beta \in (v_-, v_+)$, which will be determined later.

We divide the whole interval $\mathbf{R}$ into two subintervals $\mathbf{R}_-$ and $\mathbf{R}_+$. First, fix c and β arbitrarily, and look for solutions $(u_\pm, v_\pm)$ of the following boundary value problem on each subinterval $\mathbf{R}_\pm$ with the aid of outer and inner approximations:

$$\begin{aligned} &\varepsilon^2(u_\pm)_{xx} - \varepsilon c\tau(u_\pm)_x + f(u_\pm, v_\pm) = 0, \\ &(v_\pm)_{xx} - c(v_\pm)_x + g(u_\pm, v_\pm) = 0, \\ &u_\pm(\pm\infty) = u_\pm, \qquad u_\pm(0) = \alpha, \\ &v_\pm(\pm\infty) = v_\pm, \qquad v_\pm(0) = \beta. \end{aligned} \qquad x \in \mathbf{R}_\pm, \tag{2.5}_\pm$$

Second, we derive two relations between c and β through C^1-matching of the outer and inner solutions of $(2.5)_\pm$ at $x = 0$, respectively, and construct singular limit solutions by taking the intersection of these two relations. Finally, using a singular limit solution, we obtain a solution (u, v) of (2.1), (2.2) for some $c = c(\varepsilon)$ (see Fig. 3).

2.1. Outer solutions. Fig. 3 shows that the derivatives of $u_\pm(x)$ are moderate in the region away from a layer position. Therefore, the solutions of the following limiting equations of $(2.5)_\pm$ as $\varepsilon \downarrow 0$ could become good approximations there:

$$\begin{aligned} &f(u_\pm, v_\pm) = 0, \\ &(v_\pm)_{xx} - c(v_\pm)_x + g(u_\pm, v_\pm) = 0, \qquad x \in \mathbf{R}_\pm, \\ &v_\pm(\pm\infty) = v_\pm, \qquad v_\pm(0) = \beta. \end{aligned} \tag{2.6}_\pm$$

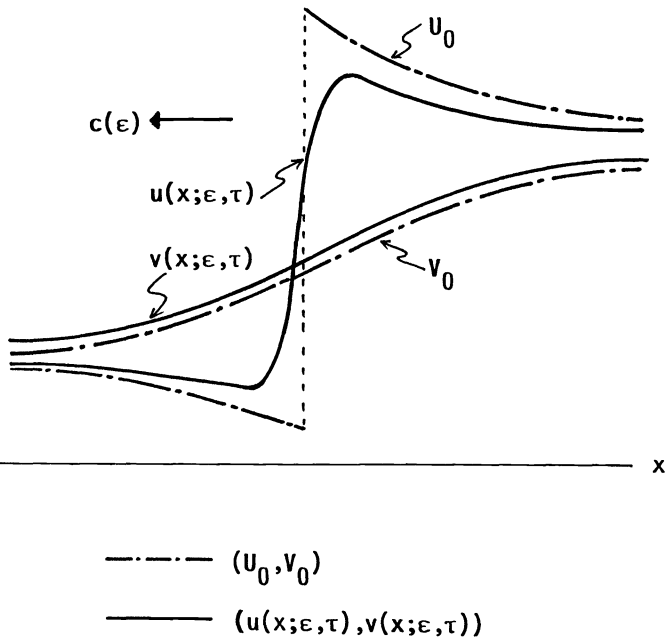


FIG. 3. *The profile of a traveling front solution* $(u(x; \varepsilon, \tau), v(x; \varepsilon, \tau))$ *and its outer solution* (U_0, V_0).

As particular solutions of the first equation, we take $u_\pm = h_\pm(v_\pm)$ (see (A1)). Substituting this into the second equation, we see that $(2.6)_\pm$ is reduced to

$$(2.7)_\pm \qquad \begin{aligned} &(V_\pm)_{xx} - c(V_\pm)_x + g(h_\pm(V_\pm), V_\pm) = 0, \qquad x \in \mathbf{R}_\pm, \\ &V_\pm(\pm\infty) = v_\pm, \qquad V_\pm(0) = \beta. \end{aligned}$$

LEMMA 2.1. *For any fixed $c \in \mathbf{R}$ and $\beta \in (v_-, v_+)$, there exist unique strictly monotone increasing solutions $V_0^\pm(x; c, \beta)$ of $(2.7)_\pm$ satisfying*

$$|V_0^\pm(x; c, \beta) - v_\pm| \in X^2_{\mu(c),1}(\mathbf{R}_\pm)$$

where $\mu(c) = \min\{\mu_-(c), \mu_+(c)\}$ and $\mu_\pm(c)$ are positive roots of $\mu_\pm^2 - c\mu_\pm + (d/dv)g(h_\pm(v_\pm), v_\pm) = 0$. Moreover, $V_0^\pm(x; c, \beta)$ are continuous with respect to $(c, \beta) \in \mathbf{R} \times (v_-, v_+)$ in the $X^2_{\mu(c),1}(\mathbf{R}_\pm)$-topology and satisfy

$$(2.8) \qquad \frac{\partial}{\partial c}\left[\frac{d}{dx} V_0^-(0; c, \beta) - \frac{d}{dx} V_0^+(0; c, \beta)\right] > 0$$

and

$$(2.9) \qquad \frac{\partial}{\partial \beta}\left[\frac{d}{dx} V_0^-(0; c, \beta) - \frac{d}{dx} V_0^+(0; c, \beta)\right] > 0.$$

LEMMA 2.2. (outer matching condition). *For any fixed $c \in \mathbf{R}$, there uniquely exists $\beta = \beta_0(c) \in C^1(\mathbf{R})$ satisfying*

$$\frac{d}{dx} V_0^-(0; c, \beta_0(c)) - \frac{d}{dx} V_0^+(0; c, \beta_0(c)) = 0,$$

which is a strictly monotone decreasing function of $c \in \mathbf{R}$ and converges to $v_\pm$ as $c \to \mp\infty$, respectively. Moreover, for $v^* \in (v_-, v_+)$,

$$I(v^*) \gtreqless 0 \quad \text{if and only if} \quad \beta_0(0) \gtreqless v^*$$

where $I(\beta) = \int_{v_-}^{\beta} g(h_-(v), v)\, dv + \int_{\beta}^{v_+} g(h_+(v), v)\, dv$.

We define $U_0^\pm(x; c, \beta)$ by

$$U_0^\pm(x; c, \beta) = h_\pm(V_0^\pm(x; c, \beta)), \qquad x \in \mathbf{R}_\pm.$$

We denote the C^1-matching *outer solution* on the whole line by

$$(2.10)_a \qquad V_0(x; c) \equiv \begin{cases} V_0^-(x; c, \beta_0(c)), & x \in \mathbf{R}_-, \\ V_0^+(x; c, \beta_0(c)), & x \in \mathbf{R}_+ \end{cases}$$

and

$$(2.10)_b \qquad U_0(x; c) \equiv \begin{cases} h_-(V_0^-(x; c, \beta_0(c))), & x \in \mathbf{R}_-, \\ h_+(V_0^+(x; c, \beta_0(c))), & x \in \mathbf{R}_+ \end{cases}$$

(see Fig. 3.)

2.2. Inner solutions. Since the outer solutions $U_0^\pm(x; c, \beta)$ do not satisfy the boundary condition at $x = 0$, we must remedy them in a neighborhood of $x = 0$. For this purpose, it is convenient to introduce the stretched variable $y = x/\varepsilon$. Substituting $(U_0^\pm + W_0^\pm, V_0^\pm)$ into $(2.5)_\pm$ with remedy terms $W_0^\pm(y)$, and putting $\varepsilon = 0$, we obtain the following problems for $W_0^\pm$:

$$(2.11)_\pm \qquad \begin{aligned} &(W_0^\pm)_{yy} - c\tau(W_0^\pm)_y + f(h_\pm(\beta) + W_0^\pm, \beta) = 0, \qquad y \in \mathbf{R}_\pm, \\ &W_0^\pm(0) = \alpha - h_\pm(\beta), \\ &W_0^\pm(\pm\infty) = 0 \end{aligned}$$

where β and α are arbitrarily fixed constants satisfying $\beta \in (v_-, v_+)$ and $\alpha \in (h_-(\beta), h_+(\beta))$. That is, the inner transition layer is stretched on the whole line and connects $h_-(\beta)$ to $h_+(\beta)$.

LEMMA 2.3 (Fife and McLeod [6]). *For any fixed $\beta \in [v_-, v_+]$, consider the following problem*:

$$\begin{aligned} &W_{yy} - cW_y + f(W, \beta) = 0, \qquad y \in \mathbf{R}, \\ &W(\pm\infty) = h_\pm(\beta), \qquad W(0) = \alpha. \end{aligned} \tag{2.12}$$

Then there exists $c = c_0(\beta)$ such that (2.12) *has a unique strictly monotone increasing solution $W(y; c_0(\beta), \beta)$ satisfying*

$$|W(y; c_0(\beta); \beta) - h_\pm(\beta)| \in X^2_{\sigma_\pm(\beta),1}(\mathbf{R}_\pm)$$

where

$$\sigma_\pm(\beta) = [\mp c_0(\beta) + \sqrt{(c_0(\beta))^2 - 4f_u(h_\pm(\beta), \beta)}]/2$$

and

$$c_0(\beta) \gtreqless 0 \quad \textit{if and only if } J(\beta) \gtreqless 0.$$

It is almost clear from Lemma 2.3 that the derivatives of W_0^+ and W_0^- are matched at $x = 0$ if and only if c is equal to $c_0(\beta)/\tau$.

LEMMA 2.4 (inner matching condition). *For any fixed $\beta \in [v_-, v_+]$, let*

$$c_I(\beta; \tau) \equiv c_0(\beta)/\tau. \tag{2.13}$$

Then there exists $\delta_0 > 0$ such that for any fixed $(\hat{c}, \hat{\beta}) \in \Lambda_{\delta_0} \equiv \{(\hat{c}, \hat{\beta}) \,|\, |\hat{c} - c_I(\beta; \tau)| + |\hat{\beta} - \beta| \leqq \delta_0\}$, $(2.11)_\pm$ have unique strictly monotone increasing solutions $W_0^\pm(y; \tau, \hat{c}, \hat{\beta})$ satisfying

$$|W_0^\pm(y; \tau, \hat{c}, \hat{\beta}) - h_\pm(\hat{\beta})| \in X^2_{\sigma_\pm(\tau),1}(\mathbf{R}_\pm)$$

where $\sigma_\pm(\tau) = \inf_{(\hat{c},\hat{\beta}) \in \Lambda_{\delta_0}} \sigma_\pm(\tau; \hat{c}, \hat{\beta})$ with

$$\sigma_\pm(\tau; c, \beta) \equiv [\mp c\tau + \sqrt{(c\tau)^2 - 4f_u(h_\pm(\beta), \beta)}]/2.$$

Furthermore, $W_0^\pm(y; \tau, \hat{c}, \hat{\beta})$ are continuous with respect to $(\hat{c}, \hat{\beta}) \in \Lambda_{\delta_0}$ in the $X^2_{\sigma_\pm(\tau),1}(\mathbf{R}_\pm)$-topology and

$$\frac{d}{dy} W_0^-(0; \tau, c_I(\beta; \tau), \beta) - \frac{d}{dy} W_0^+(0; \tau, c_I(\beta; \tau), \beta) = 0, \tag{2.14}$$

$$\frac{\partial}{\partial c}\left[\frac{d}{dy} W_0^-(0; \tau, c_I(\beta; \tau), \beta) - \frac{d}{dy} W_0^+(0; \tau, c_I(\beta; \tau), \beta)\right] > 0, \tag{2.15}_a$$

$$\frac{\partial}{\partial \beta}\left[\frac{d}{dy} W_0^-(0; \tau, c_I(\beta; \tau), \beta) - \frac{d}{dy} W_0^+(0; \tau, c_I(\beta; \tau), \beta)\right] > 0. \tag{2.15}_b$$

Remark 2.1. It follows from (2.14) and (2.15) that $(d/d\beta)c_I(\beta, \tau) (= (1/\tau) \times (d/d\beta)c_0(\beta))$ is strictly negative for $\beta \in [v_-, v_+]$. Therefore there exists an inverse function of (2.13), say $\beta = \beta_I(c; \tau)$, that is strictly decreasing for $c \in (c_I(v_+; \tau), c_I(v_-; \tau))$.

Remark 2.2. The definition domain for β can be extended to $(v_{\min}, v_{\max})$ in Lemma 2.4, since Lemma 2.3 holds for $\beta \in (v_{\min}, v_{\max})$.

2.3. Singular limit traveling front solutions. It is clear that the lowest-order approximations $(U_0^\pm(x; c, \beta) + W_0^\pm(x; \tau, c, \beta), V_0^\pm(x; c, \beta))$ of $(2.5)_\pm$ are matched at $x = 0$ in the C^0-sense. To construct an exact solution for small positive ε on the whole

line $\mathbf{R}$, the singular perturbation method requests C^1-continuity of these approximations so that their derivatives must be matched at $x=0$ in the C^0-sense. Thus, we impose the following conditions on $(W_0^{\pm}, V_0^{\pm})$:

$$\begin{aligned}&\Phi_0(\tau, c, \beta) \equiv \frac{d}{dy} W_0^-(0; \tau, c, \beta) - \frac{d}{dy} W_0^+(0; \tau, c, \beta) = 0,\\ &\Psi_0(c, \beta) \equiv \frac{d}{dx} V_0^-(0; c, \beta) - \frac{d}{dx} V_0^+(0; c, \beta) = 0.\end{aligned} \tag{2.16}$$

It turns out from Lemmas 2.2 and 2.4 that the above relations are equivalent to the conditions:

$$\beta = \beta_0(c) \tag{2.17}$$

and

$$c = c_0(\beta)/\tau. \tag{2.18}$$

From Remark 2.1, the latter is equivalent to

$$\beta = \beta_I(c; \tau). \tag{2.19}$$

Note that both β_0 and β_I are C^1-functions and strictly decreasing as in Fig. 4 (see also Remark 2.2). Geometrically, the solution set for (2.16) is represented by the intersection points of two curves $\beta = \beta_0(c)$ and $\beta = \beta_I(c; \tau)$.

For any given $\tau > 0$, let (c^*, β^*) be an arbitrary intersection point of (2.17) and (2.19). Define $(u_0(x; \varepsilon, \tau), v_0(x; \varepsilon, \tau))$ by

$$u_0(x; \varepsilon, \tau) = \begin{cases} U_0^-(x; c^*, \beta^*) + W_0^-\left(\dfrac{x}{\varepsilon}; \tau, c^*, \beta^*\right), & x \in \mathbf{R}_-, \\ U_0^+(x; c^*, \beta^*) + W_0^+\left(\dfrac{x}{\varepsilon}; \tau, c^*, \beta^*\right), & x \in \mathbf{R}_+ \end{cases} \tag{2.20}_a$$

and

$$v_0(x; \varepsilon, \tau) = \begin{cases} V_0^-(x; c^*, \beta^*), & x \in \mathbf{R}_-, \\ V_0^+(x; c^*, \beta^*), & x \in \mathbf{R}_+. \end{cases} \tag{2.20}_b$$

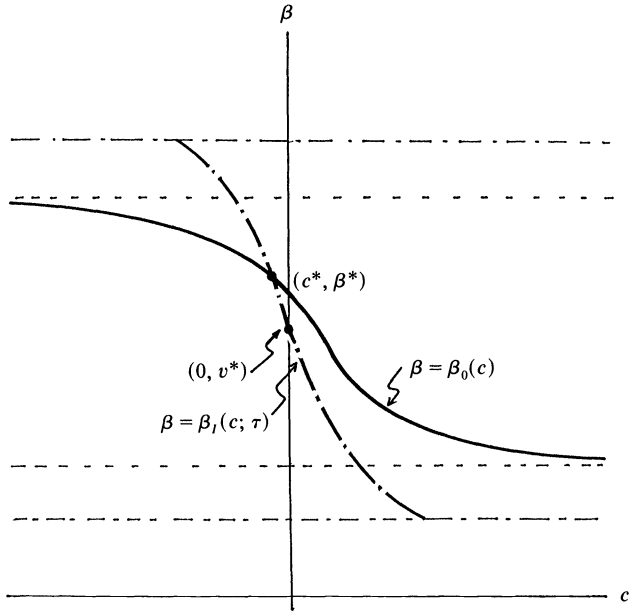
We call $(u_0(x; \varepsilon, \tau), v_0(x; \varepsilon, \tau))$ a *singular limit traveling front solution* of (2.1)-(2.3) with the *singular limit velocity* c^*.

The following theorem shows that the number of the singular limit traveling front solutions varies depending on τ and the location of v^* (see (A2)).

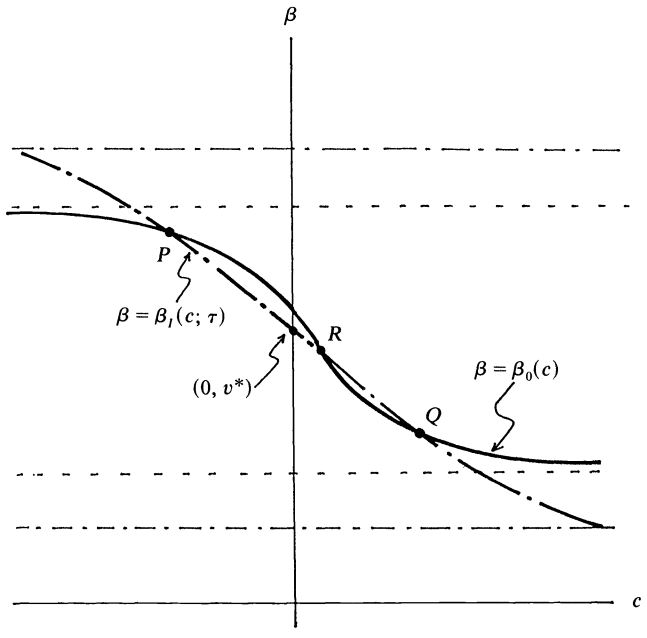
THEOREM 2.1. *Suppose that* (A0)-(A4) *hold. When* $v^* \in (v_-, v_+)$, (2.1)-(2.3) *has three singular limit traveling front solutions for small* τ *and has only one for large* τ. (*See Fig.* 2.) *On the other hand, when* $v^* \in (v_{\min}, v_{\max}) \backslash (v_-, v_+)$, *it has only one for both small and large* τ.

2.4. Traveling front solutions for $\varepsilon > 0$. Using a singular limit traveling front solution as an approximation, we can construct the exact solution of (2.1)-(2.3) with the aid of the standard singular perturbation method. Let (c^*, β^*) be an arbitrary intersection point of (2.17) and (2.19), and assume that

$$\frac{\partial(\Phi_0, \Psi_0)}{\partial(c, \beta)} \neq 0 \quad \text{at } (c^*, \beta^*), \tag{2.21}$$



(a) τ: large



(b) τ: small

FIG. 4. *The graphs of outer and inner matching conditions and their intersections. For large τ they have a unique intersection, but for small τ they have three intersecting points.*

which is equivalent to the following:

(2.22) The two curves of C^1-matching conditions $\beta=\beta_0(c)$ and $\beta=\beta_I(c;\tau)$ intersect with each other *transversally* at (c^*,β^*).

First we fix τ, c, and β and construct exact solutions of $(2.5)_\pm$ on each subinterval $\mathbf{R}_\pm$ for sufficiently small ε, which we denote by $(u^\pm(x;\varepsilon,\tau,c,\beta), v^\pm(x;\varepsilon,\tau,c,\beta))$ (see Lemma 3.2 of [10]), and then match these solutions at $x=0$ in C^1-sense. For this purpose, we define two functions Φ and Ψ as follows:

$$\begin{aligned}\Phi(\varepsilon,\tau,c,\beta)&\equiv\varepsilon\frac{d}{dx}u^-(0;\varepsilon,\tau,c,\beta)-\varepsilon\frac{d}{dx}u^+(0;\varepsilon,\tau,c,\beta),\\ \Psi(\varepsilon,\tau,c,\beta)&\equiv\frac{d}{dx}v^-(0;\varepsilon,\tau,c,\beta)-\frac{d}{dx}v^+(0;\varepsilon,\tau,c,\beta),\end{aligned}\tag{2.23}$$

and determine c and β as functions of ε such that

$$\Phi(\varepsilon,\tau,c,\beta)=0=\Psi(\varepsilon,\tau,c,\beta)\tag{2.24}$$

hold. Noting that Φ, Ψ, and their first derivatives with respect to c and β are uniformly continuous for $\varepsilon>0$, we can extend them continuously up to $\varepsilon=0$. Letting $\varepsilon=0$, (2.23) is reduced to (2.16), i.e.,

$$\Phi(0,\tau,c,\beta)=\Phi_0(\tau,c,\beta),\qquad \Psi(0,\tau,c,\beta)=\Psi_0(c,\beta).\tag{2.25}$$

As before, for any fixed $\tau>0$, let (c^*,β^*) be an arbitrary solution of (2.25). Recalling the nondegenerate condition (2.21), we easily see that

$$\frac{\partial(\Phi,\Psi)}{\partial(c,\beta)}\neq 0\tag{2.26}$$

holds at $(\varepsilon,\tau,c,\beta)=(0,\tau,c^*,\beta^*)$. Thus we can apply the Implicit Function Theorem to (2.24). That is, there is $\varepsilon_0>0$ such that there exist continuous functions $c(\varepsilon;\tau)$ and $\beta(\varepsilon;\tau)$ satisfying (2.24) for $\varepsilon\in[0,\varepsilon_0)$ and $\lim_{\varepsilon\downarrow 0}c(\varepsilon;\tau)=c^*$ and $\lim_{\varepsilon\downarrow 0}\beta(\varepsilon;\tau)=\beta^*$. We have reached the goal.

THEOREM 2.2. *Suppose that* (A0)-(A4) *hold and that, for a given* $\tau>0$, *the curves* (2.17) *and* (2.19) *intersect transversally at* (c^*,β^*). *Then, for any* $\varepsilon\in(0,\varepsilon_0)$ *there exists a traveling front solution* $(u(x;\varepsilon,\tau),v(x;\varepsilon,\tau))\in X^2_{\rho,\varepsilon}(\mathbf{R})\times X^2_{\rho,1}(\mathbf{R})$ *of the problem* (2.1)-(2.3), *satisfying*

$$\|u(\cdot;\varepsilon,\tau)-u_0(\cdot;\varepsilon,\tau)\|_{X^1_{\rho,\varepsilon}(\mathbf{R})}+\|v(\cdot;\varepsilon,\tau)-v_0(\cdot;\varepsilon,\tau)\|_{X^1_{\rho,1}(\mathbf{R})}\to 0$$

as $\varepsilon\downarrow 0$. *Furthermore, the velocity* $c(\varepsilon;\tau)$ *converges to the singular velocity* c^* *as* $\varepsilon\downarrow 0$.

We simply denote this solution by $\mathcal{U}^\varepsilon=(u^\varepsilon,v^\varepsilon)$.

COROLLARY 2.1. *Suppose that* (A0)-(A4) *hold and fix* ε *to be sufficiently small. When* $v^*\in(v_-,v_+)$, (2.1)-(2.3) *has three traveling front solutions for small* τ *and has only one for large* τ. *On the other hand, when* $v^*\in(v_{\min},v_{\max})\backslash(v_-,v_+)$, *it has only one for both small and large* τ.

Let us specify f and g as (1.1). Then the solution structure is revealed for all τ as in Fig. 2.

Finally, we show the asymptotic behavior of the stretched traveling front solutions of Theorem 2.2 on any compact interval as $\varepsilon\downarrow 0$, which plays an important role in the next section.

LEMMA 2.5. *Let* $(u^\varepsilon, v^\varepsilon) = (u(x;\varepsilon,\tau), v(x;\varepsilon,\tau))$ *be a traveling front solution obtained in Theorem* 2.2, *and let* $(\tilde{u}^\varepsilon, \tilde{v}^\varepsilon)$ *be the stretched solution of* $(u^\varepsilon, v^\varepsilon)$, *namely,* $(\tilde{u}^\varepsilon, \tilde{v}^\varepsilon) \equiv (u(\varepsilon y;\varepsilon,\tau), v(\varepsilon y;\varepsilon,\tau))$. *Then we have*

$$\lim_{\varepsilon \downarrow 0} (\tilde{u}^\varepsilon, \tilde{v}^\varepsilon) = (W(y; c_0(\beta^*), \beta^*), \beta^*) \quad \text{in } C^2_{\text{c.u.}}(\mathbf{R})\text{-sense} \tag{2.27}$$

where $W(y; c_0(\beta^*), \beta^*)$ *is the unique monotone increasing solution of* (2.12).

Proof. By using (2.20) and Theorem 2.2, we can easily show (2.27). So we leave the details to the reader. See also Lemma 1.1 of [13].

Remark 2.3. Note that (2.27) contains the following result:

$$\lim_{\varepsilon \downarrow 0} \frac{d}{dy} \tilde{u}^\varepsilon = \frac{d}{dy} W(y; c_0(\beta^*), \beta^*) \quad \text{in } C^1_{\text{c.u.}}(\mathbf{R})\text{-sense.}$$

The following result is a direct consequence of Lemma 2.5.

COROLLARY 2.2. *Let* $F(u, v)$ *be a smooth function of* u *and* v. *Then, the composite function* $F(\tilde{u}^\varepsilon, \tilde{v}^\varepsilon)$ *satisfies*

$$\lim_{\varepsilon \downarrow 0} F(\tilde{u}^\varepsilon, v^\varepsilon) = F(W(y; c_0(\beta^*), \beta^*), \beta^*) \quad \text{in } C^2_{\text{c.u.}}(\mathbf{R})\text{-sense.}$$

We close this section by presenting a lemma on the embedding properties of the Hilbert space $H^1_\rho(\mathbf{R})$.

LEMMA 2.6. *The Hilbert space* $H^1_\rho(\mathbf{R})$ $(\rho > 0)$ *satisfies the following properties*:

(i) $H^1_\rho(\mathbf{R})$ *is continuously embedded in* $C_{\text{unif}}(\mathbf{R})$.

(ii) *A bounded set in* $H^1_\rho(\mathbf{R})$ $(\rho > 0)$ *is precompact in* $L^2_{\rho'}(\mathbf{R})$ *for* $0 < \rho' < \rho$.

Proof. Taking an expanding sequence of compact intervals that converges to the whole line, and applying the diagonal arguments with the Sobolev Embedding Theorem on a compact interval, we can prove two claims without difficulty by virtue of the exponential weight function in the definition of $H^1_\rho(\mathbf{R})$.

3. Criterion for the stability of traveling front solutions. In this section, we will study the stability of the traveling fronts obtained in the previous section. If we consider the linearized equations of $(P)_{\varepsilon,\tau}$ around the specified traveling front solutions, the spectrum of the resulting linearized problem consists of two parts; the essential spectrum and isolated eigenvalues. It is proved later that the former one is not dangerous (see Proposition 3.1) but the latter one is crucial to the stability (see Lemma 3.9 and Theorem 3.2). In § 3.3 we clarify the limiting location of the real isolated eigenvalues as $\varepsilon \downarrow 0$, which varies depending on the parameter τ. Finally, we show in § 3.4 that the limiting analysis in § 3.3 is valid for small but positive ε.

3.1. Linearized problem and preliminaries. Let us take an arbitrary traveling front solution $\mathcal{U}^\varepsilon \equiv (u^\varepsilon, v^\varepsilon)$ of (2.1)-(2.3). Recall that when τ is arbitrarily fixed, the velocity c is determined as a function of ε with the singular limit velocity $c^* = \lim_{\varepsilon \downarrow 0} c(\varepsilon)$. The original evolutional system $(P)_{\varepsilon,\tau}$ takes the following form after using the traveling coordinate $x = z + c(\varepsilon)t$ and shifting the origin to $\mathcal{U}^\varepsilon$:

$$\begin{pmatrix} \varepsilon\tau & 0 \\ 0 & 1 \end{pmatrix} \frac{\partial}{\partial t} \begin{pmatrix} \hat{u} \\ \hat{v} \end{pmatrix} = \mathcal{L}^\varepsilon \begin{pmatrix} \hat{u} \\ \hat{v} \end{pmatrix} + H(\hat{u}, \hat{v}) \tag{$\hat{P}$}_{\varepsilon,\tau}$$

where

$$\begin{pmatrix} \hat{u} \\ \hat{v} \end{pmatrix} \equiv \begin{pmatrix} u \\ v \end{pmatrix} - \begin{pmatrix} u^\varepsilon \\ v^\varepsilon \end{pmatrix}, \qquad \mathcal{L}^\varepsilon \equiv \begin{pmatrix} L^\varepsilon & f^\varepsilon_v \\ g^\varepsilon_u & M^\varepsilon \end{pmatrix}$$

with

$$L^\varepsilon \equiv \varepsilon^2 \frac{d^2}{dx^2} - c(\varepsilon)\varepsilon\tau \frac{d}{dx} + f_u^\varepsilon, \qquad M^\varepsilon \equiv \frac{d^2}{dx^2} - c(\varepsilon)\frac{d}{dx} + g_v^\varepsilon,$$

and

$$H(\hat{u}, \hat{v}) \equiv \begin{pmatrix} f(u, v) - f(u^\varepsilon, v^\varepsilon) - f_u^\varepsilon \hat{u} - f_v^\varepsilon \hat{v} \\ g(u, v) - g(u^\varepsilon, v^\varepsilon) - g_u^\varepsilon \hat{u} - g_v^\varepsilon \hat{v} \end{pmatrix}.$$

Here $f_u^\varepsilon, f_v^\varepsilon, g_u^\varepsilon$, and g_v^ε denote, respectively, the partial derivatives of f and g evaluated at $\mathcal{U}^\varepsilon$. The corresponding linearized eigenvalue problem is given by

$$(LP)_{\varepsilon,\tau} \qquad \mathcal{L}^\varepsilon \begin{pmatrix} w \\ z \end{pmatrix} \equiv \begin{bmatrix} L^\varepsilon & f_v^\varepsilon \\ g_u^\varepsilon & M^\varepsilon \end{bmatrix}, \qquad \begin{pmatrix} w \\ z \end{pmatrix} = \lambda \begin{pmatrix} \varepsilon\tau & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} w \\ z \end{pmatrix}.$$

The underlying space for $(\hat{P})_{\varepsilon,\tau}$ and $(LP)_{\varepsilon,\tau}$ can be taken as $\mathbf{X}_1 \equiv \mathbf{C}_{\text{unif}}(\mathbf{R})$ with

$$D(\mathcal{L}^\varepsilon) = \{\mathbf{U} = (w, z)^t \mid \mathbf{U}, \mathbf{U}_x, \mathbf{U}_{xx} \in \mathbf{X}_1\}. \tag{3.1}$$

Using the standard arguments, we can show that $\mathcal{L}^\varepsilon$ becomes a sectorial operator (see, for example, Henry [8]) and the spectral distribution of $\mathcal{L}^\varepsilon$ determines the nonlinear stability (or instability) in $\mathbf{X}_1^\alpha$-topology, where $\mathbf{X}_1^\alpha$ denotes the Banach space associated with the fractional power of $\mathcal{L}^\varepsilon$, namely,

$$\mathbf{X}_1^\alpha \equiv D((-\mathcal{L}^\varepsilon + \kappa I)^\alpha) \tag{3.2}$$

for an appropriate positive constant κ and $\alpha \in [0, 1)$ with the usual graph norm. Although the choice of the underlying space might be a problem of taste depending on the phenomena described by the model systems, $\mathbf{C}_{\text{unif}}(\mathbf{R})$ seems to have a natural topology for reaction-diffusion systems. Another choice is, for example, $L^2(\mathbf{R})$; however, in this case, the initial perturbation must satisfy some sort of decaying property at infinity. Nevertheless, it should be noted that there are no essential differences among underlying spaces as far as the discrete spectrum of $(LP)_{\varepsilon,\tau}$ is concerned, since the associated eigenfunctions decay exponentially as $|x| \to \infty$. Taking this advantage, we will look for the eigenfunctions in $H_\rho^1(R)$ $(\subset \mathbf{L}^2(\mathbf{R}))$ instead of $\mathbf{C}_{\text{unif}}(\mathbf{R})$, which is more convenient for our purposes. In view of $(LP)_{\varepsilon,\tau}$, we see that it becomes highly degenerated as $\varepsilon \downarrow 0$: the highest order of L^ε vanishes, and each coefficient in $\mathcal{L}^\varepsilon$ has a discontinuous point at the layer position in this limit. Therefore, it is not clear in advance what kind of singular behaviors and degeneracies will occur for eigenvalues as well as eigenfunctions in the limit of $\varepsilon \downarrow 0$. As we will see in § 3.3, the SLEP method is very useful in solving these problems, and it has worked nicely in various other problems (see Nishiura and Fujii [13], [14], Nishiura [12], and Nishiura and Mimura [15]).

In the remaining part of this section, we will present several preliminaries used to derive the SLEP equation in the next section. Hereafter we simply write c instead of $c(\varepsilon)$. The main thing is to clarify the spectral behavior of the singular Sturm–Liouville eigenvalue problem:

$$\begin{aligned} &L^\varepsilon \phi = \zeta \phi, \\ &\phi \in D(L^\varepsilon) \equiv \{\phi \mid \phi, \phi_x, \phi_{xx} \in C_{\text{unif}}(\mathbf{R})\}. \end{aligned} \tag{3.3}$$

Let ζ_0^ε and ϕ_0^ε be the principal eigenvalue and its eigenfunction of (3.3) satisfying

$\|\phi_0^\varepsilon\|_{L^2(\mathbf{R})} = 1$, respectively. It is convenient to introduce the stretched problem of (3.3):

$$\tilde{L}^\varepsilon \tilde{\phi} \equiv \left(\frac{d^2}{dy^2} - c\tau\frac{d}{dy} + \tilde{f}_u^\varepsilon\right)\tilde{\phi} = \zeta\tilde{\phi}, \tag{3.4}$$

$$\tilde{\phi} \in D(\tilde{L}^\varepsilon) \equiv \{\tilde{\phi} \mid \tilde{\phi}, \tilde{\phi}_y, \tilde{\phi}_{yy} \in C_{\text{unif}}(\mathbf{R})\}$$

where $y \equiv x/\varepsilon$ is a stretched variable and $\tilde{f}_u^\varepsilon$ is defined by $\tilde{f}_u^\varepsilon \equiv f_u(\tilde{u}^\varepsilon, \tilde{v}^\varepsilon)$. The L^2-normalized principal eigenfunction $\hat{\phi}_0^\varepsilon$ of (3.4) is defined by

$$\hat{\phi}_0^\varepsilon \equiv \sqrt{\varepsilon}\ \tilde{\phi}_0^\varepsilon = \sqrt{\varepsilon}\ \phi_0^\varepsilon(\varepsilon y). \tag{3.5}$$

Of course, the eigenvalues for (3.4) remain the same by stretching.

Remark 3.1. In general, the existence of the discrete spectrum for (3.3) is not trivial, however, we will show in the proof of Lemma 3.2 that there exists the principal eigenvalue of (3.3), which tends to zero as $\varepsilon \downarrow 0$, and that the associated eigenfunction decays exponentially as $|x| \to +\infty$.

Recalling Corollary 2.2, we see that the limiting form of the potential term $\tilde{f}_u^\varepsilon$ becomes

$$\lim_{\varepsilon \downarrow 0} \tilde{f}_u^\varepsilon = f_u(W(y; c_0(\beta^*), \beta^*), \beta^*) \quad \text{in } C^2_{\text{c.u.}}\text{-sense.}$$

We denote this limiting function by $\tilde{f}_u^*$. Therefore, the limiting Sturm–Liouville problem of (3.4) becomes

$$\tilde{L}^*\tilde{\phi} \equiv \left(\frac{d^2}{dy^2} - c^*\tau\frac{d}{dy} + \tilde{f}_u^*\right)\tilde{\phi} = \zeta\phi, \tag{3.6}$$

$$\tilde{\phi} \in D(\tilde{L}^*) = D(\tilde{L}^\varepsilon).$$

Remark 3.2. Differentiating (2.12) with respect to y for $c = c_0(\beta^*)$ $(= c^*\tau)$, we see that the limiting problem (3.6) has zero eigenvalue and the associated eigenfunction is given by $W_y(y; c_0(\beta^*), \beta^*)$.

Formally, the adjoint problem for (3.3) is given by

$$(L^\varepsilon)^* \phi^* = \zeta^* \phi^*, \qquad \phi^* \in D(L^\varepsilon)^* = D(L^\varepsilon) \tag{3.7}$$

where

$$(L^\varepsilon)^* \equiv \varepsilon^2 \frac{d^2}{dx^2} + c\varepsilon\tau\frac{d}{dx} + f_u^\varepsilon.$$

Apparently, the stretched adjoint problem is defined by

$$(\tilde{L}^\varepsilon)^* \tilde{\phi}^* = \zeta^* \tilde{\phi}^* \tag{3.8}$$

where

$$(\tilde{L}^\varepsilon)^* \equiv \frac{d^2}{dy^2} + c\tau\frac{d}{dy} + \tilde{f}_u^\varepsilon$$

with the same definition domain as (3.4). We denote by $\phi_0^{\varepsilon *}$ (respectively, $\hat{\phi}_0^{\varepsilon *} \equiv \sqrt{\varepsilon}\ \phi_0^{\varepsilon *}(\varepsilon y)$) the L^2-normalized principal eigenfunction of (3.7) (respectively, (3.8)) associated with the principal eigenvalue $\zeta^* = \zeta_0^\varepsilon$. Applying the change of dependent variables from ϕ (respectively, ϕ^*) to $\psi \equiv e^{-(c\tau/2\varepsilon)x}\phi$ (respectively, $\psi \equiv e^{(c\tau/2\varepsilon)x}\phi^*$), it is easily seen that (3.3) (respectively, (3.7)) is converted to the formal self-adjoint operator:

$$L_s^\varepsilon \psi \equiv \left[\varepsilon^2 \frac{d^2}{dx^2} + \left\{f_u^\varepsilon - \left(\frac{c\tau}{2}\right)^2\right\}\right]\psi = \zeta\psi. \tag{3.9}$$

Remark 3.3. Since the element in $D(L^\varepsilon)$ has no decaying property near $\pm\infty$, the formal adjoint operator of L^ε is not defined as (3.7). Therefore, at the present stage, it is not appropriate to call (3.7) the adjoint problem for (3.3). However, as we will see in the proof of Lemma 3.2, the eigenfunctions in $C_{\text{unif}}(\mathbf{R})$ of (3.3) (or (3.4)) decay exponentially as $|x|\to\infty$ (or $|y|\to\infty$), which may justify the above abuse. In fact, we will see that the problem (3.3) (or (3.7)) is equivalent to (3.9) in $L^2(\mathbf{R})$, as far as isolated eigenvalues are concerned.

Remark 3.4. There exists a one-to-one correspondence between the isolated eigenvalues and their eigenfunctions of L^ε and those of $(L^\varepsilon)^\#$. Namely, suppose (ζ, ϕ) (respectively, $(\zeta, \tilde{\phi})$) is an eigenpair of (3.3) (respectively, (3.4)); then $(\zeta, e^{-c\tau x/\varepsilon}\phi)$ (respectively, $(\zeta, e^{-c\tau y}\tilde{\phi})$) becomes an eigenpair of (3.7) (respectively, (3.8)). Note that eigenvalues are unchanged.

The principal eigenfunction of (3.4) and the y-derivative of the stretched u-component of the traveling front solution converge to the stretched inner layer solution as follows.

LEMMA 3.1. *Let ϕ_0^ε (respectively, $\phi_0^{\varepsilon *}$) be the L^2-normalized principal eigenfunction of L^ε (respectively, $(L^\varepsilon)^\#$), and let $\hat{\phi}_0^\varepsilon$ (respectively, $\hat{\phi}_0^{\varepsilon *}$) be the L^2-normalized stretched function of ϕ_0^ε (respectively, $\phi_0^{\varepsilon *}$), namely, $\hat{\phi}_0^\varepsilon = \sqrt{\varepsilon}\,\phi_0^\varepsilon(\varepsilon y)$ (respectively, $\hat{\phi}_0^{\varepsilon *} = \sqrt{\varepsilon}\,\phi_0^{\varepsilon *}(\varepsilon y)$). Then it holds that*

(i) $\lim_{\varepsilon\downarrow 0} \hat{\phi}_0^\varepsilon$ *(respectively,* $\tilde{u}_y^\varepsilon$*)* $= \gamma W_y(y; c_0(\beta^*), \beta^*)$ *(respectively,* $W_y(y; c_0(\beta^*), \beta^*)$*) in* $C^2_{\text{c.u.}}(\mathbf{R})$*-sense,*

(ii) $\lim_{\varepsilon\downarrow 0} \hat{\phi}_0^{\varepsilon *} = \gamma^\# W_y^\#(y; c_0(\beta^*), \beta^*)$ *in* $C^2_{\text{c.u.}}(\mathbf{R})$*-sense*

where γ and $\gamma^\#$ are the positive normalized constants given by

$$\gamma = \| W_y(y; c_0(\beta^*), \beta^*)\|_{L^2}^{-1} \quad \text{and} \quad \gamma^\# = \| W_y^\#(y; c_0(\beta^*), \beta^*)\|_{L^2}^{-1},$$

respectively, with $W_y^\#(y; c_0(\beta^*), \beta^*) \equiv e^{-c_0(\beta^*)\tau y} W_y(y; c_0(\beta^*), \beta^*)$.

Proof. In view of the construction of $(u^\varepsilon, v^\varepsilon)$ (see § 2, Ikeda, Mimura, and Nishiura [10]), and Remark 3.4, we can prove the above lemma as in Lemma 1.3 of Nishiura and Fujii [13].

LEMMA 3.2 (spectral properties of L^ε). *The essential spectrum of* (3.3) *is contained in the union of the left regions inside or at the boundaries of the two parabolas*:

$$\text{Re}\,\zeta = -(\text{Im}\,\zeta)^2/(\tau c)^2 + \alpha_\pm \tag{3.10}$$

where $\alpha_\pm = \lim_{x\to\pm\infty} f_u^\varepsilon < 0$. *The spectrum lying outside the above region consists of real isolated eigenvalues, and they have a strictly negative upper bound $-\mu_0$ for small ε except the principal eigenvalue ζ_0^ε, where μ_0 is a positive constant independent of ε. The principal eigenvalue ζ_0^ε is the unique critical eigenvalue of* (3.3) *(i.e., it approaches zero as $\varepsilon\downarrow 0$) and behaves as*

$$\zeta_0^\varepsilon = \hat{\zeta}_0(\varepsilon)\varepsilon \quad \text{as } \varepsilon\downarrow 0 \tag{3.11}$$

where $\hat{\zeta}_0(\varepsilon)$ is a continuous function of ε up to $\varepsilon = 0$ satisfying

$$\hat{\zeta}_0^* \equiv \lim_{\varepsilon\downarrow 0} \hat{\zeta}_0(\varepsilon) = -\left\{ c^*(\beta^* - v_-) - \int_{-\infty}^{0} g(U_0, V_0)\,dx \right\} \frac{dc_0(\beta^*)}{d\beta} > 0 \tag{3.12}$$

where (U_0, V_0) is the outer solution defined in (2.10) *and $c_0(\beta)$ is the inner velocity defined in Lemma* 2.3. (See Fig. 5.)

Proof. (i) Location of the essential spectrum. Since the traveling front solutions $\mathcal{U}^\varepsilon$ converge to the critical point with exponential order as $|x|\to\infty$, the coefficient f_u^ε of (3.3) becomes an asymptotically negative constant as $|x|\to\infty$. It is known (see, for example, Henry [8]) that the location of the essential spectrum $\sigma_e(L^\varepsilon)$ for such an

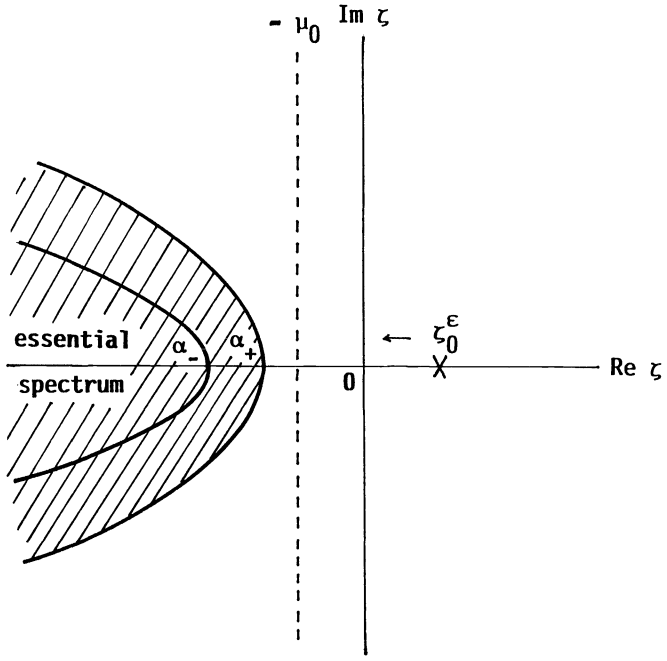


FIG. 5. *Spectral behavior of the singular Sturm–Liouville operator* L^ε.

operator as L^ε is contained in the union of the left regions inside or the boundaries of the following curves:

$$S_\pm = \{\zeta \mid -\varepsilon^2\gamma^2 - ic\varepsilon\tau\gamma + \alpha_\pm - \zeta = 0,\ -\infty < \gamma < \infty\} \tag{3.13}$$

where $\lim_{x\to\pm\infty} f_u^\varepsilon = \alpha_\pm$. It is easy to see that the sets $S_\pm$ are two parabolas in $\mathbf{C}$ defined by

$$\operatorname{Re}\zeta = -(\operatorname{Im}\zeta)^2/(c\tau)^2 + \alpha_\pm. \tag{3.14}$$

Note that the above two parabolas are uniformly bounded away from the imaginary axis for small ε.

(ii) Location of the isolated eigenvalues. Let ζ lie outside the union of the left regions inside or the boundaries of $S_\pm$. Suppose that such a ζ belongs to the spectrum of (3.3); then it must be an isolated eigenvalue. First, we will show that the associated eigenfunction must decay exponentially as $|x| \to \infty$. To do this, it suffices to consider the limiting first-order systems of (3.3) as $x \to \pm\infty$:

$$\frac{d}{dx}\begin{pmatrix}\phi\\ \phi_x\end{pmatrix} = \begin{pmatrix}0 & 1/\varepsilon\\ (-\alpha_\pm+\zeta)/\varepsilon & c\tau/\varepsilon\end{pmatrix}\begin{pmatrix}\phi\\ \phi_x\end{pmatrix}. \tag{3.15}_\pm$$

The eigenvalues of the matrix of the right-hand side are given by

$$\frac{c\tau}{2\varepsilon} \pm \frac{1}{2}\sqrt{\left(\frac{c\tau}{\varepsilon}\right)^2 + \frac{4}{\varepsilon^2}(-\alpha_\pm+\zeta)}. \tag{3.16}$$

Here we denote by $\frac{1}{2}\sqrt{(c\tau/\varepsilon)^2 + (4/\varepsilon^2)(-\alpha_\pm+\zeta)}$ the complex number whose real part is greater than $|c|\tau/2\varepsilon$. Note that the numbers in (3.16) become pure imaginary numbers if and only if ζ lies on the parabolic curves $S_\pm$ (see (3.13)). Since we assume that ζ is outside the parabolic regions, the real parts of (3.16) are not equal to zero. Therefore,

the real parts of two eigenvalues of (3.16) have the opposite sign. Suppose ζ is an eigenvalue and ϕ is the associated eigenfunction of (3.3) in $C_{\text{unif}}(\mathbf{R})$. Then it is well known from the general theory (see, for example, Coddington and Levinson [1]) that ϕ has the same asymptotic behaviors as those of solutions of $(3.15)_{\pm}$ when $x \to \pm\infty$. Namely, ϕ must decay exponentially as

$$|\phi| \simeq \exp\left\{\frac{c\tau}{2\varepsilon} \mp r_{\pm}(\zeta)\right\} x \quad \text{as } x \to \pm\infty \tag{3.17}$$

where $r_{\pm}(\zeta)$ are defined by

$$r_{\pm}(\zeta) \equiv \frac{1}{2} \operatorname{Re} \sqrt{\left(\frac{c\tau}{\varepsilon}\right)^2 + \frac{4}{\varepsilon^2}(-\alpha_{\pm} + \zeta)} \qquad \left(> \frac{|c|\tau}{2\varepsilon}\right). \tag{3.18}$$

Now let us apply the change of independent variables from ϕ to ψ as

$$\psi \equiv e^{-(c\tau/2\varepsilon)x}\phi. \tag{3.19}$$

Then, (3.3) is converted to the self-adjoint form (3.9). Note that the eigenvalues are invariant. It follows from (3.17) and (3.18) that ψ also decays exponentially strictly faster than $O(\exp(-(|c|\tau/2\varepsilon)|x|))$ in both directions $x \to \pm\infty$. Conversely, it is clear that any eigenvalue of (3.9) is real and the corresponding eigenfunction that belongs to $L^2(\mathbf{R})$ must decay exponentially as $|x| \to +\infty$. Therefore, it is easily seen that there is a *one-to-one correspondence* of the isolated eigenvalues between the problem (3.3) in $C_{\text{unif}}(\mathbf{R})$ and the problem (3.9) and $L^2(\mathbf{R})$. In fact they are exactly the same. Thus we can conclude that the eigenvalues of (3.3) are real and the associated eigenfunctions decay as does (3.17).

Next we will show the existence of the principal eigenvalue ζ_0^{ε}, and that the remaining eigenvalues (if they exist) are strictly smaller than ζ_0^{ε} up to $\varepsilon = 0$. It is convenient to consider this problem in the self-adjoint form (3.9) in $L^2(\mathbf{R})$. The principal eigenvalue ζ_0^{ε} is characterized by the maximum value of the variational problem:

$$\max Q^{\varepsilon}(\psi, \psi) \equiv \max\left\{-\varepsilon^2\langle\psi_x, \psi_x\rangle + \left\langle\left\{f_u^{\varepsilon} - \left(\frac{c\tau}{2}\right)^2\right\}\psi, \psi\right\rangle\right\} \tag{3.20}$$

where ψ varies in $H^1(\mathbf{R})$ satisfying $\|\psi\|_{L^2(\mathbf{R})} = 1$.

First note that ${}^{\mathcal{T}}u_x^{\varepsilon} = (u_x^{\varepsilon}, v_x^{\varepsilon})$ satisfies $(LP)_{\varepsilon,\tau}$ with $\lambda = 0$. Therefore, applying the same change of variables as (3.19) to u_x^{ε}, we find that $\psi^{\varepsilon} \equiv e^{-(c\tau/2\varepsilon)x}u_x^{\varepsilon}$ satisfies $L_s^{\varepsilon}\psi^{\varepsilon} + f_v^{\varepsilon}\, e^{-(c\tau/2\varepsilon)x}v_x^{\varepsilon} = 0$. Thus $\bar{\psi}^{\varepsilon} \equiv \psi^{\varepsilon}/\|\psi^{\varepsilon}\|_{L^2(\mathbf{R})}$ satisfies

$$Q^{\varepsilon}(\bar{\psi}^{\varepsilon}, \bar{\psi}^{\varepsilon}) = \langle -f_v^{\varepsilon}\, e^{-(c\tau/2\varepsilon)x}v_x^{\varepsilon}/\|\psi^{\varepsilon}\|_{L^2(\mathbf{R})}, \bar{\psi}^{\varepsilon}\rangle.$$

In view of Lemma 3.1 and the construction of $(u^{\varepsilon}, v^{\varepsilon})$, we see that $\langle -f_v^{\varepsilon}\, e^{-(c\tau/2\varepsilon)x}v_x^{\varepsilon}, \bar{\psi}^{\varepsilon}\rangle$ is uniformly bounded for any small ε and that $\|\psi^{\varepsilon}\|_{L^2(\mathbf{R})} = O(1/\sqrt{\varepsilon})$. Hence, $Q^{\varepsilon}(\bar{\psi}^{\varepsilon}, \bar{\psi}^{\varepsilon})$ satisfies

$$Q^{\varepsilon}(\bar{\psi}^{\varepsilon}, \bar{\psi}^{\varepsilon}) \geqq -c\sqrt{\varepsilon}$$

for some positive constant c that is strictly larger than the supremum of the essential spectrum (see (3.10)). This implies the existence of the isolated principal eigenvalue ζ_0^{ε}. Let ψ_0^{ε} attain the maximum value; then $|\psi_0^{\varepsilon}|$ also attains the same value. Therefore we can take ψ_0^{ε} to be nonnegative. Moreover, noting that $\psi_0^{\varepsilon} \in H^2(\mathbf{R})$, ψ_0^{ε} is strictly positive (so is the principal eigenfunction ϕ_0^{ε} $(\equiv e^{(c\tau/2\varepsilon)x}\psi_0^{\varepsilon})$ of (3.3)), since if $\psi(x) = 0 = \psi'(x)$ at some point, ψ becomes identically zero. Let $\bar{\zeta}^{\varepsilon}$ be an arbitrary eigenvalue of (3.9) and let ψ^{ε} be an associated eigenfunction; then it holds that

$$\varepsilon^2 \int_{-\infty}^{\infty} (\psi_0^{\varepsilon})^2 \left\{\frac{d}{dx}\left(\frac{\psi^{\varepsilon}}{\psi_0^{\varepsilon}}\right)\right\}^2 dx = \zeta_0^{\varepsilon} - \bar{\zeta}^{\varepsilon}. \tag{3.21}$$

Apparently, $\zeta_0^\varepsilon \geqq \bar{\zeta}^\varepsilon$ holds. Suppose $\bar{\zeta}^\varepsilon = \zeta_0^\varepsilon$; then it follows from (3.21) that $\psi^\varepsilon / \psi_0^\varepsilon =$ const., which implies that ζ_0^ε is a simple eigenvalue. Now we will show that all the other eigenvalues are strictly below the principal eigenvalue ζ_0^ε up to $\varepsilon = 0$. Namely, there exists a positive constant μ_0 independent of ε such that

$$\zeta_0^\varepsilon - \bar{\zeta}^\varepsilon \geqq \mu_0 > 0 \tag{3.22}$$

holds for any eigenvalue $\bar{\zeta}^\varepsilon$ $(\neq \zeta_0^\varepsilon)$ and small ε. To show (3.22), it is convenient to introduce the stretched variable

$$y = \frac{x}{\varepsilon}. \tag{3.23}$$

Then (3.9) becomes

$$\tilde{L}_s^\varepsilon \tilde{\psi} \equiv \left[\frac{d^2}{dy^2} + \left\{ \tilde{f}_u^\varepsilon - \left(\frac{c\tau}{2} \right)^2 \right\} \right] \tilde{\psi} = \zeta \tilde{\psi}. \tag{3.24}$$

Let $\hat{\psi}_0^\varepsilon$ be defined by

$$\hat{\psi}_0^\varepsilon \equiv \sqrt{\varepsilon}\, \psi_0^\varepsilon(\varepsilon y). \tag{3.25}$$

Then $\hat{\psi}_0^\varepsilon$ is the principal eigenfunction of (3.24) with $\|\hat{\psi}_0^\varepsilon\|_{L^2(\mathbf{R})} = 1$. Similarly, $\hat{\psi}^\varepsilon$ is defined to be $\sqrt{\varepsilon}\, \psi^\varepsilon(\varepsilon y)$. Formula (3.21) can be rewritten as

$$\int_{-\infty}^{\infty} (\hat{\psi}_0^\varepsilon)^2 \left\{ \frac{d}{dy} \left(\frac{\hat{\psi}^\varepsilon}{\hat{\psi}_0^\varepsilon} \right) \right\}^2 dy = \zeta_0^\varepsilon - \bar{\zeta}^\varepsilon. \tag{3.26}$$

Suppose the claim (3.22) does not hold. Then we can find a sequence $\varepsilon_n \downarrow 0$ as $n \uparrow \infty$ such that there exists another eigenvalue $\zeta_1^{\varepsilon_n}$ satisfying

$$\lim_{n \uparrow \infty} \zeta_0^{\varepsilon_n} = \lim_{n \uparrow \infty} \zeta_1^{\varepsilon_n}. \tag{3.27}$$

Here we use the fact that ζ_0^ε remains bounded for small ε. We can assume without loss of generality that $\zeta_1^{\varepsilon_n}$ is the second eigenvalue of (3.9) (or (3.3)). Let us denote by ψ_1^ε (or $\hat{\psi}_1^\varepsilon$ in y variable) the corresponding normalized eigenfunction to ζ_1^ε. Note the following orthogonal property:

$$\langle \psi_1^\varepsilon, \psi_0^\varepsilon \rangle_x = 0 \quad \text{and} \quad \langle \hat{\psi}_1^\varepsilon, \hat{\psi}_0^\varepsilon \rangle_y = 0. \tag{3.28}$$

Here we need Sublemma 3.1.

SUBLEMMA 3.1. *There exist a subsequence* $\{\varepsilon_{n'}\}$, *of* $\{\varepsilon_n\}$, *and two functions* $\hat{\psi}_0^\infty$ *and* $\hat{\psi}_1^\infty$ *in* $H^2(\mathbf{R})$ *such that*

$$\lim_{n' \uparrow \infty} \hat{\psi}_i^{\varepsilon_{n'}} = \hat{\psi}_i^\infty \quad \text{in } H^2(\mathbf{R})\text{-sense for } i = 0, 1.$$

Proof. Noting that both eigenfunctions decay exponentially as $|x| \to \infty$, it can be shown that both families of functions remain bounded in $H_\rho^1(\mathbf{R})$ for an appropriate $\rho > 0$. Using Lemma 2.6 and (3.24), we can easily reach the conclusion. The details are left to the reader.

Substituting the results of Sublemma 3.1 and (3.27) into (3.26), we see in the limit of $\varepsilon_{n'} \downarrow 0$ that

$$\int_{-\infty}^{\infty} (\hat{\psi}_0^\infty)^2 \left\{ \frac{d}{dy} \left(\frac{\hat{\psi}_1^\infty}{\hat{\psi}_0^\infty} \right) \right\}^2 dy = 0. \tag{3.29}$$

This implies that $\hat{\psi}_1^\infty$ is a constant multiple of $\hat{\psi}_0^\infty$. However, this contradicts the orthogonal property (3.28), which is also valid for $\hat{\psi}_0^\infty$ and $\hat{\psi}_1^\infty$. This completes the proof of (3.22).

Finally, we will prove the asymptotic behavior of the principal eigenvalue ζ_0^ε when $\varepsilon \downarrow 0$. The basic fact is that zero is an eigenvalue of $\mathscr{L}^\varepsilon$ and the associated eigenfunction is given by the spatial derivative of $\mathscr{U}^\varepsilon$ as has been seen before. Namely, it holds that

$$L^\varepsilon u_x^\varepsilon + f_v^\varepsilon v_x^\varepsilon = 0, \qquad g_u^\varepsilon u_x^\varepsilon + M^\varepsilon v_x^\varepsilon = 0. \tag{3.30}$$

Using the stretched variable y, this becomes

$$\tilde{L}^\varepsilon \tilde{u}_y^\varepsilon + \tilde{f}_v^\varepsilon v_y^\varepsilon = 0, \qquad \tilde{g}_u^\varepsilon \tilde{u}_y^\varepsilon + \tilde{M}^\varepsilon v_y^\varepsilon = 0 \tag{3.31}$$

where

$$\tilde{L}^\varepsilon = \frac{d^2}{dy^2} - c\tau \frac{d}{dy} + \tilde{f}_u^\varepsilon, \quad \tilde{M}^\varepsilon = \frac{1}{\varepsilon^2}\frac{d^2}{dy^2} - \frac{c}{\varepsilon}\frac{d}{dy} + \tilde{g}_v^\varepsilon, \quad \tilde{u}_y^\varepsilon = \frac{d}{dy} u^\varepsilon(\varepsilon y),$$

and so on. Recalling the construction of $\mathscr{U}^\varepsilon = (u^\varepsilon, v^\varepsilon)$ (see § 2), we see that both $\tilde{u}_y^\varepsilon$ and $\tilde{v}_y^\varepsilon$ decay exponentially as $|y| \to \infty$. The adjoint operator $(\tilde{L}^\varepsilon)^* \equiv d^2/dy^2 + c\tau(d/dy) + \tilde{f}_u^\varepsilon$ has the same principal eigenvalue ζ_0^ε and the corresponding eigenfunction $\hat{\phi}_0^{\varepsilon *}$ is given by $e^{-c\tau y}\hat{\phi}_0^\varepsilon$ (see Remark 3.4). Multiplying $\hat{\phi}_0^{\varepsilon *}$ to the first equation of (3.31), after integration by parts we obtain

$$\langle \tilde{u}_y^\varepsilon, (\tilde{L}^\varepsilon)^* \hat{\phi}_0^{\varepsilon *} \rangle + \langle \tilde{f}_v^\varepsilon \tilde{v}_y^\varepsilon, \hat{\phi}_0^{\varepsilon *} \rangle = 0.$$

Since $(\tilde{L}^\varepsilon)^* \hat{\phi}_0^{\varepsilon *} = \zeta_0^\varepsilon \hat{\phi}_0^{\varepsilon *}$, we have

$$\zeta_0^\varepsilon = \frac{\langle -\tilde{f}_v^\varepsilon \tilde{v}_y^\varepsilon, \hat{\phi}_0^{\varepsilon *} \rangle}{\langle \tilde{u}_y^\varepsilon, \hat{\phi}_0^{\varepsilon *} \rangle}. \tag{3.32}$$

On the other hand, $\tilde{v}^\varepsilon$ satisfies

$$\frac{1}{\varepsilon^2}\frac{d^2}{dy^2}\tilde{v}^\varepsilon - \frac{c}{\varepsilon}\frac{d}{dy}\tilde{v}^\varepsilon + g(\tilde{u}^\varepsilon, \tilde{v}^\varepsilon) = 0.$$

Integrating this with respect to y, we have

$$\tilde{v}_y^\varepsilon = \varepsilon \left\{ c(\tilde{v}^\varepsilon - \tilde{v}^\varepsilon(-\infty)) - \varepsilon \int_{-\infty}^{y} g(\tilde{u}^\varepsilon, \tilde{v}^\varepsilon)\, dy \right\}. \tag{3.33}$$

Substituting (3.33) into (3.32), we obtain

$$\zeta_0^\varepsilon = \varepsilon \left\langle -\tilde{f}_v^\varepsilon \left\{ c(\tilde{v}^\varepsilon - v_-) - \varepsilon \int_{-\infty}^{y} g(\tilde{u}^\varepsilon, \tilde{v}^\varepsilon)\, dy \right\}, \hat{\phi}_0^{\varepsilon *} \right\rangle \Big/ \langle \tilde{u}_y^\varepsilon, \hat{\phi}_0^{\varepsilon *} \rangle. \tag{3.34}$$

By using Lemma 3.1 and Corollary 2.2, the numerator and the denominator of (3.34) have the following limits as $\varepsilon \downarrow 0$:

$$\begin{aligned} &\lim_{\varepsilon \downarrow 0} \left\langle \left\{ c(\tilde{v}^\varepsilon - v_-) - \varepsilon \int_{-\infty}^{y} g(\tilde{u}^\varepsilon, \tilde{v}^\varepsilon)\, dy \right\} (-\tilde{f}_v^\varepsilon), \hat{\phi}_0^{\varepsilon *} \right\rangle \\ &= \left\{ c^*(V_0(0) - v_-) - \int_{-\infty}^{0} g(U_0, V_0)\, dx \right\} \cdot \lim_{\varepsilon \downarrow 0} \langle -\tilde{f}_v^\varepsilon, \hat{\phi}_0^{\varepsilon *} \rangle \qquad (3.35) \\ &= \left\{ c^*(\beta^* - v_-) - \int_{-\infty}^{0} g(U_0, V_0)\, dx \right\} \langle -f_v(W, \beta^*), \gamma^* W_y^* \rangle \end{aligned}$$

and

$$\lim_{\varepsilon \downarrow 0} \langle \tilde{u}_y^\varepsilon, \hat{\phi}_0^{\varepsilon *} \rangle = \langle W_y, \gamma^* W_y^* \rangle. \tag{3.36}$$

Recall that $W(y; c_0(\beta), \beta)$ satisfies (see Lemma 2.3)

$$\frac{d^2}{dy^2} W - c_0(\beta) \frac{d}{dy} W + f(W, \beta) = 0.$$

Differentiating this with respect to β, we have

$$\frac{d^2}{dy^2} W_\beta - c_0(\beta) \frac{d}{dy} W_\beta + f_u(W, \beta) W_\beta - \frac{d}{d\beta} c_0(\beta) \frac{d}{dy} W + f_v(W, \beta) = 0. \tag{3.37}$$

Taking the inner product with $W_y^\#$ on both sides of (3.37) and using the fact that $W_y^\#$ satisfies

$$\frac{d^2}{dy^2} W_y^\# + c_0(\beta) \frac{d}{dy} W_y^\# + f_u(W, \beta) W_y^\# = 0,$$

we obtain for $\beta = \beta^*$

$$-\frac{d}{d\beta} c_0(\beta) \langle W_y, W_y^\# \rangle + \langle f_v(W, \beta^*), W_y^\# \rangle = 0. \tag{3.38}$$

Substituting (3.35), (3.36), and (3.38) into (3.34), we can conclude that

$$\lim_{\varepsilon \downarrow 0} \frac{\zeta_0^\varepsilon}{\varepsilon} = -\frac{d}{d\beta} c_0(\beta) \left\{ c(\beta^* - v_-) - \int_{-\infty}^{0} g(U_0, V_0)\, dx \right\},$$

which is strictly positive from Remark 2.1 and the strict monotonicity of V_0, i.e.,

$$\frac{d}{dy} V_0(0) = c^*(\beta^* - v_-) - \int_{-\infty}^{0} g(U_0, V_0)\, dx > 0.$$

This completes the proof of Lemma 3.2.

3.2. Location of the essential spectrum. In this section we will consider the location of the essential spectrum of $(LP)_{\varepsilon,\tau}$. It is known that there are several different kinds of definitions for the essential spectrum. Here we adopt the definition employed in Goldberg [7] (see also Henry [8]). However, it should be noted that even if we take a different one, we can obtain the same a priori bound as in Proposition 3.1, since it is known that $\sup \sigma_e\{(LP)_{\varepsilon,\tau}\}$ does not depend on the choice of the definitions (see, for example, Edmunds and Evans [2]). Our goal is the following.

PROPOSITION 3.1. *For a given $\tau > 0$, there exists a positive constant δ_e independent of ε and τ such that*

$$\operatorname{Re}\{\sigma_e(LP)_{\varepsilon,\tau}\} \leqq -\delta_e < 0$$

holds for small ε.

Proof. The location of the essential spectrum is determined by the following sets:

$$S_\pm = \{\lambda \mid \det(-\mu^2 D - i\mu M + N_\pm - \lambda B) = 0, -\infty < \mu < \infty\} \tag{3.39}$$

where

$$D = \begin{pmatrix} \varepsilon^2 & 0 \\ 0 & 1 \end{pmatrix}, \qquad M = \begin{pmatrix} c\varepsilon\tau & 0 \\ 0 & c \end{pmatrix}$$

$$N_\pm = \begin{pmatrix} \alpha_\pm & \beta_\pm \\ \gamma_\pm & \delta_\pm \end{pmatrix} \equiv \lim_{x \to \pm\infty} \begin{pmatrix} f_u^\varepsilon & f_v^\varepsilon \\ g_u^\varepsilon & g_v^\varepsilon \end{pmatrix}, \qquad B = \begin{pmatrix} \varepsilon\tau & 0 \\ 0 & 1 \end{pmatrix}.$$

It is clear that

$$\det(-\mu^2 D - i\mu M + N_\pm - \lambda B) = \begin{vmatrix} -\mu^2\varepsilon^2 - i\mu c\varepsilon\tau + \alpha_\pm - \varepsilon\tau\lambda & \beta_\pm \\ \gamma_\pm & -\mu^2 - i\mu c + \delta_\pm - \lambda \end{vmatrix} = 0$$

becomes

$$\varepsilon\tau\lambda^2 - \{-(\varepsilon\tau+\varepsilon^2)\mu^2 + (\alpha_\pm + \varepsilon\tau\delta_\pm) - 2i\mu c\varepsilon\tau\}\lambda + (\mu^2\varepsilon^2 + i\mu c\varepsilon\tau - \alpha_\pm)(\mu^2 + i\mu c - \delta_\pm) - \beta_\pm\gamma_\pm = 0.$$

The roots of this equation are given by

$$\lambda = \frac{1}{2\varepsilon\tau}[-(\varepsilon\tau+\varepsilon^2)\mu^2 + (\alpha_\pm + \varepsilon\tau\delta_\pm) - 2i\mu c\varepsilon\tau \pm \sqrt{\Theta}]$$

where

$$\Theta \equiv \{(\varepsilon\tau+\varepsilon^2)\mu^2 - (\alpha_\pm + \varepsilon\tau\delta_\pm)\}^2 - 4(\mu c\varepsilon\tau)^2 + 4i\mu c\varepsilon\tau\{(\varepsilon\tau+\varepsilon^2)\mu^2 - (\alpha_\pm + \varepsilon\tau\delta_\pm)\} - 4\varepsilon\tau\{(\mu^2\varepsilon^2 - \alpha_\pm + i\mu c\varepsilon\tau)(\mu^2 - \delta_\pm + i\mu c) - \beta_\pm\gamma_\pm\}.$$

After some computation, we have

$$\operatorname{Re}\Theta = \{(\varepsilon\tau-\varepsilon^2)\mu^2 + (\alpha_\pm - \varepsilon\tau\delta_\pm)\}^2 + 4\varepsilon\tau\beta_\pm\gamma_\pm, \qquad \operatorname{Im}\Theta = 0.$$

Therefore,

$$\lambda = \frac{1}{2\varepsilon\tau}[-(\varepsilon\tau+\varepsilon^2)\mu^2 + (\alpha_\pm + \varepsilon\tau\delta_\pm) - 2i\mu c\varepsilon\tau \pm \sqrt{\{(\varepsilon\tau-\varepsilon^2)\mu^2 + (\alpha_\pm - \varepsilon\tau\delta_\pm)\}^2 + 4\varepsilon\tau\beta_\pm\gamma_\pm}]. \tag{3.40}$$

We will compute the supremum of the real part of (3.40) when μ varies in $\mathbf{R}$, which gives us the upper bound of the essential spectrum. Suppose the inside of the $\sqrt{\ }$-part of (3.40) takes the minus sign or zero; then it is clear that

$$\operatorname{Re}\lambda \leqq \frac{1}{2\varepsilon\tau}[-(\varepsilon\tau+\varepsilon^2)\mu^2 + (\alpha_\pm + \varepsilon\tau\delta_\pm)]. \tag{3.41}$$

It is obvious that the right-hand side of (3.41) is majorized by

$$\frac{1}{2\varepsilon\tau}\alpha_\pm + \frac{1}{2}\delta_\pm, \tag{3.42}$$

which is apparently strictly negative uniformly for small ε and $\tau > 0$ from (A3). On the other hand, if the inside of the $\sqrt{\ }$-part takes the plus sign, we need to take it into account. Without loss of generality, it suffices to consider the case where the $\sqrt{\ }$-part becomes a real number, namely,

$$\{(\varepsilon\tau-\varepsilon^2)\mu^2 + (\alpha_\pm - \varepsilon\tau\delta_\pm)\}^2 + 4\varepsilon\tau\beta_\pm\gamma_\pm \quad (\geqq 0).$$

In this case, it follows from $\beta_\pm\gamma_\pm < 0$ (see (A4)) that

$$\begin{aligned}\operatorname{Re}\lambda &= \frac{1}{2\varepsilon\tau}[-(\varepsilon\tau+\varepsilon^2)\mu^2 + (\alpha_\pm + \varepsilon\tau\delta_\pm) \pm \sqrt{\{(\varepsilon\tau-\varepsilon^2)\mu^2 + (\alpha_\pm - \varepsilon\tau\delta_\pm)\}^2 + 4\varepsilon\tau\beta_\pm\gamma_\pm}] \\ &\leqq \frac{1}{2\varepsilon\tau}[-(\varepsilon\tau+\varepsilon^2)\mu^2 + (\alpha_\pm + \varepsilon\tau\delta_\pm) + |(\varepsilon\tau-\varepsilon^2)\mu^2 + (\alpha_\pm - \varepsilon\tau\delta_\pm)|].\end{aligned} \tag{3.43}$$

According to the sign of the inside of $|\cdot|$, (3.43) becomes

$$\text{(3.44)}\qquad \operatorname{Re}\lambda \leqq \begin{cases} \dfrac{1}{2\varepsilon\tau}\{-2\varepsilon^2\mu^2+2\alpha_\pm\} = -\dfrac{\varepsilon}{\tau}\mu^2+\dfrac{1}{\varepsilon\tau}\alpha_\pm \leqq \dfrac{1}{\varepsilon\tau}\alpha_\pm, \\ \dfrac{1}{2\varepsilon\tau}\{-2\varepsilon\tau\mu^2+2\varepsilon\tau\delta_\pm\} = -\mu^2+\delta_\pm \leqq \delta_\pm. \end{cases}$$

Thus we see from (3.44) that

$$\text{(3.45)}\qquad \operatorname{Re}\lambda \leqq \max\left\{\frac{1}{\varepsilon\tau}\alpha_\pm, \delta_\pm\right\}.$$

Combining the results (3.42) and (3.45), we can derive the conclusion of Proposition 3.1. Note that we can take $-\delta_e$ to be $\max\{\alpha_\pm, \frac{1}{2}\delta_\pm\}$ when ε is assumed to be taken smaller than τ^{-1} (i.e., $\varepsilon\tau<1$).

3.3. The SLEP equation and the behavior of the isolated eigenvalues in the singular limit. We will study the location of the isolated eigenvalues and their dependency on τ. As has been shown in § 3.2, the essential spectrum of $\mathscr{L}^\varepsilon$ is strictly bounded away from the imaginary axis, so that the stability properties of the traveling front solutions depend solely on the behavior of the isolated eigenvalues. In fact as we will see later, some of the isolated eigenvalues really cross the imaginary axis, and therefore we must track their behavior rather than try to obtain a priori bounds for them. The main difficulty in doing so is that, when $\varepsilon \downarrow 0$, the eigenfunctions associated with those dangerous eigenvalues do not remain in the usual function space such as $\mathbf{C}_{\text{unif}}(\mathbf{R})$ or $\mathbf{L}^2(\mathbf{R})$. The eigenfunctions actually fall into the measure space (a point measure for the one-dimensional case). However, the SLEP method enables us to overcome this difficulty and control these eigenvalues uniformly for small ε.

First, we will study the asymptotic behavior of the eigenfunctions as $|x|\to+\infty$. Since the traveling front solution $\mathscr{U}^\varepsilon$ approaches the equilibrium states E_+ and E_- with the exponential order as $|x|\to\infty$, the asymptotic behavior of the eigenfunctions for large $|x|$ can be described by the limiting constant coefficient system of $(LP)_{\varepsilon,\tau}$:

$$\text{(3.46)}_\pm\qquad \begin{aligned} &\left(\varepsilon^2\frac{d^2}{dx^2}-c\varepsilon\tau\frac{d}{dx}+\alpha_\pm\right)w+\beta_\pm z = \varepsilon\tau\lambda w, \\ &\gamma_\pm w+\left(\frac{d^2}{dx^2}-c\frac{d}{dx}+\delta_\pm\right)z = \lambda z \end{aligned}$$

where $\alpha_\pm = \lim_{x\to\pm\infty} f_u^\varepsilon = f_u(E_\pm)$, $\beta_\pm = f_v(E_\pm)$, $\gamma_\pm = g_u(E_\pm)$, and $\delta_\pm = g_v(E_\pm)$. It follows from (A3) that

$$\text{(3.47)}\qquad \alpha_\pm<0 \quad\text{and}\quad \det_\pm \equiv \alpha_\pm\delta_\pm-\beta_\pm\gamma_\pm>0.$$

It suffices to consider the + case only and, for notational simplicity, we omit the subscript + hereafter. We rewrite $(3.46)_\pm$ in the form

$$\text{(3.48)}_a\qquad \begin{aligned} &\varepsilon\frac{dw}{dx}=\xi, \qquad \varepsilon\frac{d\xi}{dx}=c\tau\xi-\alpha w-\beta z+\varepsilon\tau\lambda w, \\ &\frac{dz}{dx}=\eta, \qquad \frac{d\eta}{dx}=c\eta-\gamma w-\delta z+\lambda z \end{aligned}$$

or in a vector form

$$\frac{d}{dx}\begin{pmatrix} w \\ \xi \\ z \\ \eta \end{pmatrix} = \begin{pmatrix} 0 & 1/\varepsilon & 0 & 0 \\ -\alpha/\varepsilon + \tau\lambda & c\tau/\varepsilon & -\beta/\varepsilon & 0 \\ 0 & 0 & 0 & 1 \\ -\gamma & 0 & -\delta+\lambda & c \end{pmatrix}\begin{pmatrix} w \\ \xi \\ z \\ \eta \end{pmatrix} \tag{3.48}_b$$

$$\equiv M_\lambda^{\varepsilon,\tau,c}\begin{pmatrix} w \\ \xi \\ z \\ \eta \end{pmatrix}.$$

The eigenvalues of the matrix of $M_\lambda^{\varepsilon,\tau,c}$ determine the asymptotic behavior of the eigenfunctions. The characteristic polynomial with the unknown κ becomes

$$\kappa^2\left(\kappa - \frac{c\tau}{\varepsilon}\right)(\kappa - c) - \kappa\left(\kappa - \frac{c\tau}{\varepsilon}\right)(\lambda - \delta) - \kappa(\kappa - c)\left(\tau\lambda - \frac{\alpha}{\varepsilon}\right)\Big/\varepsilon - \frac{\beta\gamma}{\varepsilon^2} + \left(\tau\lambda - \frac{\alpha}{\varepsilon}\right)(\lambda - \delta)/\varepsilon = 0. \tag{3.49}$$

Examining the roots of (3.49), we have Proposition 3.2.

PROPOSITION 3.2. *Assume that* λ *belongs to*

$$\{\lambda \in \mathbf{C} \mid \operatorname{Re}\lambda > \det/\alpha\}. \tag{3.50}$$

The eigenvalues of the matrix $M_\lambda^{\varepsilon,\tau,c}$ *are divided into two classes. One is of the order* $O(1)$ *and the other is of the order* $O(1/\varepsilon)$ *as* $\varepsilon \downarrow 0$, *each of which consists of two eigenvalues with positive and negative* Re*-parts, respectively*:

(i) *The* $O(1)$*-class consists of two eigenvalues, say* $\kappa_\pm^1$, *which remain finite in the limit* $\varepsilon \downarrow 0$. *Their principal parts are given by*

$$\kappa_\pm^1 \equiv \frac{c}{2} \pm \sqrt{\frac{c^2}{4} - \frac{\det}{\alpha} + \lambda} + O(\varepsilon) \tag{3.51}$$

with $\operatorname{Re}\kappa_\pm^1 \gtrless 0$.

(ii) *The* $O(1/\varepsilon)$*-class has two eigenvalues, say* $\kappa_\pm^\varepsilon$, *which diverge with order* $1/\varepsilon$ *as* $\varepsilon \downarrow 0$. *The principal parts of them are given by*

$$\kappa_\pm^\varepsilon \equiv \tfrac{1}{2}(c\tau \pm \sqrt{(c\tau)^2 - 4\alpha})/\varepsilon + O(1) \tag{3.52}$$

with $\operatorname{Re}\kappa_\pm^\varepsilon \gtrless 0$.

The eigenvectors associated with the above eigenvalues are denoted by $\Xi_\pm^1$ and $\Xi_\pm^\varepsilon$, respectively. If we need to distinguish the eigenvalues and the eigenvectors at E_+ and E_-, we add + or − after the superscript one or ε such as $\kappa_\pm^{1+}$, $\Xi_\pm^{\varepsilon-}$.

Remark 3.5. Let λ satisfy

$$\operatorname{Re}\lambda > -\hat{\mu} \equiv \max\left\{\frac{\det_+}{\alpha_+}, \frac{\det_-}{\alpha_-}\right\}; \tag{3.53}$$

then, it is clear from the above proposition that, at both E_+ and E_-, (3.49) has two eigenvalues with positive real parts of the orders $O(1)$ and $O(1/\varepsilon)$ and two eigenvalues with negative real parts of the same property.

Proof of Proposition 3.2. (i) Multiplying ε^2 to (3.49) gives

$$\alpha\kappa^2 - c\alpha\kappa + (\alpha\delta - \beta\gamma) - \alpha\lambda + O(\varepsilon) = 0. \tag{3.54}$$

The two roots of the principal part of (3.54) are both simple under (3.50), leading to (3.51) with the aid of the Implicit Function Theorem.

(ii) Introducing the new unknown $\hat{\kappa}$ by $\kappa = \hat{\kappa}/\varepsilon$ and then multiplying (3.49) by ε^4, we have

$$\hat{\kappa}^2(\hat{\kappa}^2 - c\tau\hat{\kappa} + \alpha) + O(\varepsilon) = 0. \tag{3.55}$$

The two roots of the equation $\hat{\kappa}^2 - c\tau\hat{\kappa} + \alpha = 0$ are simple. Therefore, by using the Implicit Function Theorem again, we obtain (3.52). Since (3.49) is of fourth order, the above four solutions are all roots of it. The sign properties of the real parts, i.e., $\operatorname{Re}\kappa_{\pm}^1 \gtrless 0$ and $\operatorname{Re}\kappa_{\pm}^{\varepsilon} \gtrless 0$, can be shown by a simple computation under the assumption (3.50).

Proposition 3.2 clearly shows that if we suppose there is an isolated eigenvalue satisfying (3.53) with the eigenfunction Ψ in $\mathbf{C}_{\mathrm{unif}}(\mathbf{R})$, then Ψ *must* decay with the exponential order as $|x| \to \infty$. This guarantees us that later we will be able to work in much more comfortable Hilbert space $H_{\rho}^1(\mathbf{R})$ for the isolated eigenvalues.

Now we return to the problem $(LP)_{\varepsilon,\tau}$:

$$L^{\varepsilon}w + f_v^{\varepsilon}z = \varepsilon\tau\lambda w, \qquad g_u^{\varepsilon}w + M^{\varepsilon}z = \lambda z \tag{$(LP)_{\varepsilon,\tau}$}$$

where λ is an isolated eigenvalue and ${}^t(w, z)$ is the associated eigenfunction. In view of Proposition 3.2, we may think that it seems to be appropriate to regard the mapping of the left-hand side of $(LP)_{\varepsilon,\tau}$ to be the one from $X_{\rho,\varepsilon}^2(\mathbf{R}) \times X_{\rho,1}^2(\mathbf{R})$ to $X_{\rho,1}^0(\mathbf{R}) \times X_{\rho,1}^0(\mathbf{R})$. However, our basic strategy is to find the nice limiting eigenvalue problem from which the necessary information for positive ε can be easily extracted. Unfortunately, the above setting for function spaces is not fit for this purpose, since it turns out that eigenfunctions associated with the isolated eigenvalues are no longer usual functions in the limit $\varepsilon \downarrow 0$. More precisely, their w-components approach the Dirac point measure after some scaling, and, at the same time, z-components tend to smooth functions except the jump discontinuity of the first x-derivative at the layer position. Therefore, it is necessary to replace the above function spaces by weaker ones with the decaying property. For this purpose, we first convert $(LP)_{\varepsilon,\tau}$ into a single equation with respect to z by solving the first equation of $(LP)_{\varepsilon,\tau}$ with respect to w, and then we put a new appropriate function space for z, namely, $H_{\rho}^1(\mathbf{R})$ which is valid up to $\varepsilon \downarrow 0$. The following lemma is needed for this procedure.

LEMMA 3.3. *For small ε, it holds that*

$$\frac{\zeta_0^{\varepsilon}}{\varepsilon\tau} \notin \sigma\{(LP)_{\varepsilon,\tau}\}.$$

Proof. We can prove this in a similar way to that of Lemma 2.1 in [13], so we leave the details to the reader.

Owing to Lemmas 3.2 and 3.3, we can solve the first equation of $(LP)_{\varepsilon,\tau}$ with respect to w for $\lambda \in \mathbf{C}_{\hat{\mu}}$ (see Remark 3.5 for the definition of $\hat{\mu}$):

$$w = (L^{\varepsilon} - \varepsilon\tau\lambda)^{-1}(-f_v^{\varepsilon}z). \tag{3.56}$$

For later use, we decompose (3.56) into two parts

$$(L^{\varepsilon} - \varepsilon\tau\lambda)^{-1}(\cdot) = \frac{1}{\zeta_0^{\varepsilon} - \varepsilon\tau\lambda} P^{\varepsilon}(\cdot) + (L^{\varepsilon} - \varepsilon\tau\lambda)^{\dagger}(\cdot) \tag{3.57}_a$$

where P^{ε} is the projection operator on the principal eigenfunction of L^{ε} defined by

$$P^{\varepsilon}(\cdot) \equiv \langle \cdot, \phi_0^{\varepsilon *} \rangle \phi_0^{\varepsilon}, \tag{3.57}_b$$

and $(L^\varepsilon - \varepsilon\tau\lambda)^\dagger$ is the remaining part of the resolvent, i.e.,

$$(L^\varepsilon - \varepsilon\tau\lambda)^\dagger(\cdot) \equiv (L^\varepsilon - \varepsilon\tau\lambda)^{-1}(\cdot) - \frac{1}{\zeta_0^\varepsilon - \varepsilon\tau\lambda} P^\varepsilon(\cdot). \tag{3.57}_c$$

Remark 3.6. In view of Lemma 3.1 and $(3.57)_c$ (see also Lemma 3.4), it is easy to see that

$$\|(L^\varepsilon - \varepsilon\tau\lambda)^\dagger\|_{\mathscr{L}(X,X)} \leqq M, \tag{3.58}$$

holds for any small ε and $\lambda \in \mathbf{C}_{\hat{\mu}}$ with M being an appropriate positive constant independent of ε. Here the underlying space X is $L^2_\rho(\mathbf{R})$ $(\rho \geqq 0)$ and $\mathscr{L}(X, X)$ denotes the set of bounded linear operators from X to X.

Substituting (3.56) into the second equation of $(LP)_{\varepsilon,\tau}$ we have a closed eigenvalue problem with respect to z:

$$M^\varepsilon z + \frac{g_u^\varepsilon}{\zeta_0^\varepsilon - \varepsilon\tau\lambda} P^\varepsilon(-f_v^\varepsilon z) + g_u^\varepsilon (L^\varepsilon - \varepsilon\tau\lambda)^\dagger(-f_v^\varepsilon z) = \lambda z. \tag{3.59}$$

The core of the SLEP method consists of the following two key lemmas, which characterize the asymptotic behaviors of the second and the third terms of the left-hand side of (3.59).

LEMMA 3.4 (the first key lemma). *Let $F(u, v)$ be a smooth function of u and v. Then it holds that*

$$(L^\varepsilon - \varepsilon\tau\lambda)^\dagger(F^\varepsilon h) \xrightarrow[\varepsilon \downarrow 0]{} F^* h / f_u^* \quad \textit{strongly in } L^2_\rho\textit{-sense}$$

for any function $h \in L^2_\rho(\mathbf{R}) \cap L^\infty(\mathbf{R})$, $\tau \in \mathbf{R}_+$, and $\lambda \in \mathbf{C}_{\hat{\mu}}$, where $F^\varepsilon \equiv F(u^\varepsilon, v^\varepsilon)$ and $F^* = F(U_0, V_0)$. Moreover, the convergence is uniform on a bounded set in $\mathbf{C}_{\hat{\mu}} \times H^1_{\rho_1}(\mathbf{R})$ $(\rho_1 > \rho \geqq 0)$ with respect to (λ, h).

Proof. We can prove this in the spirit of the proof of Lemma 2.2 in Nishiura and Fujii [13] with the aid of Lemma 2.6. So we leave the details to the reader.

LEMMA 3.5 (the second key lemma). *It holds that*

(a) $$\lim_{\varepsilon \downarrow 0} \frac{-f_v^\varepsilon}{\sqrt{\varepsilon}} \phi_0^{\varepsilon *} = k_1^* \delta_0 \quad \textit{in } (H^1_\rho)^\#(\mathbf{R})\textit{-sense}$$

(b) $$\lim_{\varepsilon \downarrow 0} \frac{g_u^\varepsilon}{\sqrt{\varepsilon}} \phi_0^\varepsilon = k_2^* \delta_0$$

where $\delta_0 = \delta(x)$ is the Dirac δ-function at $x = 0$, and k_i^* $(i = 1, 2)$ are positive constants given by

$$k_1^* = -\gamma^\# \langle W_y, W_y^\# \rangle \frac{d}{d\beta} c_0(\beta^*) > 0,$$

$$k_2^* = \gamma\{g(h_+(\beta^*), \beta^*) - g(h_-(\beta^*), \beta^*)\} > 0.$$

Proof. With the aid of Lemma 3.1 and (3.38), we can prove this in a way similar to that of Lemma 2.3 in Nishiura and Fujii [13], so we omit it.

Now we are ready to derive the singular limit eigenvalue problem of (3.59). First we rewrite the second term of (3.59) through ε-scaling:

$$\frac{g_u^\varepsilon}{\zeta_0^\varepsilon - \varepsilon\tau\lambda} P^\varepsilon(-f_v^\varepsilon z) = \frac{\langle z, -f_v^\varepsilon \phi_0^{\varepsilon *}/\sqrt{\varepsilon} \rangle}{\zeta_0^\varepsilon/\varepsilon - \tau\lambda} g_u^\varepsilon \phi_0^\varepsilon / \sqrt{\varepsilon}\,.$$

Using the above two key lemmas and Lemma 3.2, we see that, as $\varepsilon \downarrow 0$, (3.59) becomes

$$\frac{d^2}{dx^2} z - c^* \frac{d}{dx} z + \frac{k^*}{\hat{\zeta}_0^* - \tau\lambda} \langle z, \delta_0 \rangle \delta_0 + \frac{\det^*}{f_u^*} z = \lambda z, \qquad z \in H_\rho^1(\mathbf{R}) \tag{3.60}$$

where $k^* = k_1^* k_2^*$, $\det^* \equiv f_u^* g_v^* - f_v^* g_u^*$, $f_u^* \equiv f_u(U_0, V_0)$, and so on. Here, of course, (3.60) should be written in a weak form; however, for notational simplicity, we write it in a classical form. We call (3.60) *the* SLEP *differential equation* of $(LP)_{\varepsilon,\tau}$. The formal adjoint equation to (3.60) is given by

$$\frac{d^2}{dx^2} z^\# + c^* \frac{d}{dx} z^\# + \frac{k^*}{\hat{\zeta}_0^* - \tau\lambda} \langle z^\#, \delta_0 \rangle \delta_0 + \frac{\det^*}{f_u^*} z^\# = \lambda z^\#, \qquad z^\# \in (H_\rho^1)^\#(\mathbf{R}). \tag{3.61}$$

Without loss of generality, we can take

$$\langle z, \delta_0 \rangle = 1 \tag{3.62}$$

as a normalization for z, since if $\langle z, \delta_0 \rangle = 0$, z becomes identically zero. Under (3.62), (3.60) is equivalent to

$$\frac{d^2}{dx^2} z^\pm - c^* \frac{d}{dx} z^\pm + \frac{\det^*}{f_u^*} z^\pm = \lambda z^\pm, \qquad z^\pm \in H_\rho^1(\mathbf{R}_\pm), \tag{3.63}_a$$

$$z^+(0) = z^-(0), \tag{3.63}_b$$

$$\frac{dz^+}{dx}(0) - \frac{dz^-}{dx}(0) = -\frac{k^*}{\hat{\zeta}_0^* - \tau\lambda}. \tag{3.63}_c$$

It is clear that z^+ and z^- are smooth functions of x on $\mathbf{R}_+$ and $\mathbf{R}_-$, respectively. The task is to find λ such that the associated solutions of $(3.63)_a$ and $(3.63)_b$ satisfy the jump condition $(3.63)_c$ with respect to the first derivative.

Let us convert the SLEP differential equation (3.60) into the equivalent transcendental equation, which is much easier to deal with. We first introduce the inverse of the following differential operator $T_\lambda^{\varepsilon,\tau,c}$; $H_\rho^1(\mathbf{R}) \to (H_\rho^1)^\#(\mathbf{R})$,

$$T_\lambda^{\varepsilon,\tau,c} \equiv -\frac{d^2}{dx^2} + c\frac{d}{dx} - g_u^\varepsilon (L^\varepsilon - \varepsilon\tau\lambda)^\dagger (-f_v^\varepsilon \cdot) - g_v^\varepsilon + \lambda \tag{3.64}_a$$

through the associated bilinear form

$$\begin{aligned} B_\lambda^{\varepsilon,\tau,c}(z_1, z_2) = {} & \langle (z_1)_x, (z_2)_x \rangle + c\langle (z_1)_x, z_2 \rangle \\ & - \langle g_u^\varepsilon (L^\varepsilon - \varepsilon\tau\lambda)^\dagger (-f_v^\varepsilon z_1), z_2 \rangle + \langle (\lambda - g_v^\varepsilon) z_1, z_2 \rangle \end{aligned} \tag{3.64}_b$$

for $z_1, z_2 \in H_\rho^1(\mathbf{R})$. Recall that c is not an independent parameter but a function of ε with $c^* = \lim_{\varepsilon \to 0} c(\varepsilon)$.

LEMMA 3.6. *There exist positive constants ε_0 and ρ such that the differential operator (in the generalized sense) $T_\lambda^{\varepsilon,\tau,c}: H_\rho^1(\mathbf{R}) \to (H_\rho^1)^\#(\mathbf{R})$ is uniformly invertible for $0 \leqq \varepsilon \leqq \varepsilon_0$, τ in a bounded set in $\mathbf{R}_+$, and $\lambda \in \mathbf{C}_{\hat{\mu}}$, where $\hat{\mu}$ is a positive constant stated in Remark 3.5. We denote this inverse operator by $K_\lambda^{\varepsilon,\tau,c}: (H_\rho^1)^\#(\mathbf{R}) \to H_\rho^1(\mathbf{R})$. Moreover, $K_\lambda^{\varepsilon,\tau,c}$ depends on τ and λ analytically, and depends on ε continuously up to $\varepsilon = 0$ in operator norm sense, respectively.*

Proof. We can prove this lemma in a way parallel to that of Lemma 3.1 of Nishiura and Fujii [13], so we leave the details to the reader.

Applying the operator $K_\lambda^{\varepsilon,\tau,c}$ to (3.59), we have

$$z = \frac{\langle -f_v^\varepsilon z, \phi_0^{\varepsilon *}/\sqrt{\varepsilon}\rangle}{\zeta_0^\varepsilon/\varepsilon - \tau\lambda} K_\lambda^{\varepsilon,\tau,c}\left(g_u^\varepsilon \frac{\phi_0^\varepsilon}{\sqrt{\varepsilon}}\right), \tag{3.65}$$

which shows that z is a constant multiple of $K_\lambda^{\varepsilon,\tau,c}(g_u^\varepsilon(\phi_0^\varepsilon/\sqrt{\varepsilon}))$, namely, for a constant α

$$z = \alpha K_\lambda^{\varepsilon,\tau,c}\left(g_u^\varepsilon \frac{\phi_0^\varepsilon}{\sqrt{\varepsilon}}\right) \in H_\rho^1(\mathbf{R}). \tag{3.66}$$

Substituting (3.66) into (3.65), we see that the nontrivial solution z of (3.65) exists if and only if λ satisfies the algebraic-like equation

$$\frac{\zeta_0^\varepsilon}{\varepsilon} - \tau\lambda = \left\langle K_\lambda^{\varepsilon,\tau,c}\left(g_u^\varepsilon \frac{\phi_0^\varepsilon}{\sqrt{\varepsilon}}\right), -f_v^\varepsilon \frac{\phi_0^{\varepsilon *}}{\sqrt{\varepsilon}}\right\rangle. \tag{3.67}$$

We set

$$\mathcal{F}(\lambda;\varepsilon,\tau,c) \equiv \frac{\zeta_0^\varepsilon}{\varepsilon} - \tau\lambda - \left\langle K_\lambda^{\varepsilon,\tau,c}\left(g_u^\varepsilon \frac{\phi_0^\varepsilon}{\sqrt{\varepsilon}}\right), -f_v^\varepsilon \frac{\phi_0^{\varepsilon *}}{\sqrt{\varepsilon}}\right\rangle = 0. \tag{3.68}$$

This is the basic relation among ε, τ, c, and λ in the sense that the behaviors of isolated eigenvalues with respect to ε (including $\varepsilon = 0$), τ, and c are governed by (3.68). Recalling Lemma 3.1, the left-hand side of (3.67) is defined continuously up to $\varepsilon = 0$. On the other hand, in view of Lemmas 3.5 and 3.6, the right-hand side of (3.67) is also well defined up to $\varepsilon = 0$. Therefore (3.68) holds uniformly for small ε up to $\varepsilon = 0$. Thus, the limiting equation of (3.68) as $\varepsilon \downarrow 0$ is given by

$$\mathcal{F}(\lambda;0,\tau,c^*) \equiv \hat{\zeta}_0^* - \tau\lambda - k^*\langle K_\lambda^{*,\tau,c^*}\delta_0, \delta_0\rangle = 0. \tag{3.69}$$

We call (3.69) *the* SLEP *equation* of $(LP)_{\varepsilon,\tau}$. The great advantage of the SLEP equation is that not only the limiting location of isolated eigenvalues but also the behaviors of those for positive ε can be obtained from (3.69) by applying a *usual* Implicit Function Theorem to it at $\varepsilon = 0$. This is due to the nice limiting characterization as in Lemmas 3.4 and 3.5.

We first analyze the SLEP equation (3.69), and then we will return to (3.68) in the next section. The following lemma enables us to concentrate on the study of the behavior of *real* eigenvalues of (3.69).

LEMMA 3.7. *For a given $\tau > 0$, there exists a positive constant μ_1 (which may depend on τ) such that the* SLEP *equation* (3.69) *does not have complex isolated eigenvalues in the region $\mathbf{C}_{\mu_1}$.*

Proof. See the Appendix for the proof.

The real eigenvalues of (3.69) have a simple geometrical interpretation, namely, they are the intersection points between the straight line $S(\lambda;\tau,c^*) \equiv \hat{\zeta}_0^* - \tau\lambda$ and the curve

$$G(\lambda;\tau,c^*) \equiv k^*\langle K_\lambda^{*,\tau,c^*}\delta_0, \delta_0\rangle. \tag{3.70}$$

G satisfies the following properties as a function of λ.

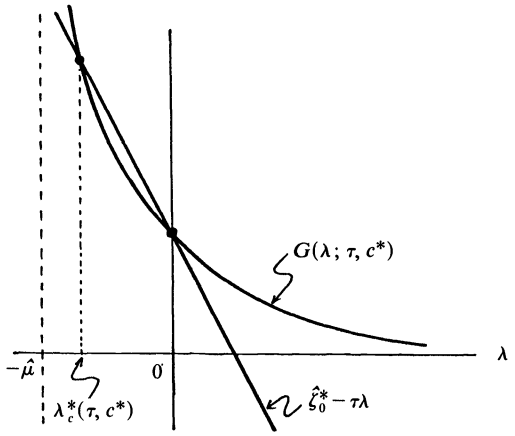
LEMMA 3.8. *$G(\lambda;\tau,c^*)$ is a strictly decreasing and convex function of real λ for $\lambda > -\hat{\mu}$ and satisfies the following*:

(i) $G(0;\tau,c^*) = \hat{\zeta}_0^*$,

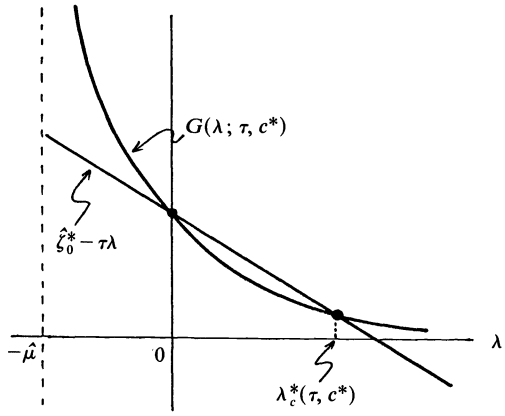
(ii) $\lim_{\lambda\to+\infty} G(\lambda;\tau,c^*) = 0$.

See Fig. 6.

Proof. We can prove this, except property (i), exactly in the same manner as the proof of Lemma 3.3 in Nishiura and Mimura [15], so we leave the details to the reader.



(a) $-\tau < \frac{d}{d\lambda} G(0; \tau, c^*)$ (stable)



(b) $-\tau > \frac{d}{d\lambda} G(0; \tau, c^*)$ (unstable)

FIG. 6. *The limiting location of the critical eigenvalue of* $(LP)_{\varepsilon,\tau}$ *is represented as an intersecting point of two curves* $G(\lambda; \tau, c^*)$ *and* $\hat{\zeta}_0^* - \tau\lambda$, *which becomes negative (respectively, positive) when* $-\tau <$ (*respectively*, $>$)$(d/d\lambda)G(0; \tau, c^*)$.

As for the property (i), this is a direct consequence of the translation invariance property. In fact, the spatial derivative of the traveling front solution $\mathcal{U}_x^\varepsilon = (u_x^\varepsilon, v_x^\varepsilon)$ is an eigenfunction of $(LP)_{\varepsilon,\tau}$ with $\lambda = 0$ for any small ε. Therefore the SLEP equation (3.69) must also have a zero eigenvalue, implying the property (i).

It is easily seen from Lemma 3.8 that there are exactly *two* intersection points including the multiplicity between S and G: as is expected, one is the zero eigenvalue

that comes from the translation invariance (see Lemma 3.8(i)), and the other is called the (real) *critical eigenvalue.* We denote this critical eigenvalue by $\lambda_c^*(\tau, c^*)$ or simply $\lambda_c^*(\tau)$. In view of Fig. 6 and Lemma 3.8, we easily see the following lemma.

LEMMA 3.9. *There exists a unique real critical eigenvalue* $\lambda = \lambda_c^*(\tau, c^*)$ *and its sign is determined as follows*:

$$\lambda_c^*(\tau, c^*) \gtreqless 0 \quad \text{if and only if} \quad -\tau \gtreqless \frac{d}{d\lambda} G(\lambda; \tau, c^*)\Big|_{\lambda=0}.$$

From Lemmas 3.7 and 3.9, and Proposition 3.1, we can conclude the following stability properties of traveling front solutions in the singular limit sense.

THEOREM 3.1 (linearized stability in the singular limit sense). *Let* $\mathcal{U}^\varepsilon = (u^\varepsilon, v^\varepsilon)$ *be a traveling front solution of* $(P)_{\varepsilon,\tau}$ *with the asymptotic velocity* c^* *as* $\varepsilon \downarrow 0$. *Then, its singular limit traveling front solution is*

$$\left.\begin{matrix} \textit{stable} \\ \textit{marginally stable} \\ \textit{unstable} \end{matrix}\right\} \quad \text{if and only if} \quad -\tau \gtreqless \frac{d}{d\lambda} G(\lambda; \tau, c^*)\Big|_{\lambda=0},$$

respectively. Here the terms stable, unstable, and marginally stable are used in the linearized sense; for example, stable means that all the spectrum have strictly negative real parts except the zero eigenvalue associated with the translation invariance.

3.4. Stability of the traveling front solutions for positive ε. In the previous section, we have studied the stability properties of the traveling front solutions in the singular limit sense. Namely, we have derived the SLEP equation (3.69), which tells us the limiting location of the real critical eigenvalue, and have given a criterion for the stability of traveling fronts (see Theorem 3.1). The great advantage of the SLEP method is that stability properties for positive ε can be obtained directly from (3.69) by using a *usual* Implicit Function Theorem, despite the fact that the linearized problem $(LP)_{\varepsilon,\tau}$ and the associated eigenfunctions behave in a singular manner as $\varepsilon \downarrow 0$.

Generically speaking, stability properties are preserved when ε becomes positive. The only exceptional case is the third one (the marginal case) in Theorem 3.1 where stability properties as well as existence for positive ε become more delicate problems. The discussions of the details for this case are postponed to a forthcoming paper where we have a nice qualitative description about what happens to the tangential case in a structurally stable manner when we look at the behavior of the solution *branch* in (τ, c)-space rather than looking at just one particular solution.

The location of eigenvalues of $(LP)_{\varepsilon,\tau}$ for positive ε is determined by (3.68), i.e.,

$$\mathcal{F}(\lambda; \varepsilon, \tau) = \frac{\zeta_0^\varepsilon}{\varepsilon} - \tau\lambda - \left\langle K_\lambda^{\varepsilon,\tau,c}\left(g_u^\varepsilon \frac{\phi_0^\varepsilon}{\sqrt{\varepsilon}}\right), -f_v^\varepsilon \frac{\phi_0^{\varepsilon *}}{\sqrt{\varepsilon}}\right\rangle = 0. \tag{3.68}$$

We can analyze this equation as a perturbation of the SLEP equation (3.69). We use the simple notation $\lambda_c^*(\tau)$ as a critical eigenvalue of (3.69) (see Lemma 3.9).

LEMMA 3.10. *The real critical eigenvalue* $\lambda_c^*(\tau)$ *of* (3.69) *can be extended uniquely to the solution* $\lambda_c^\varepsilon(\tau)$ *of* $\mathcal{F}(\lambda; \varepsilon, \tau) = 0$ *for small* ε, *where* $\lambda_c^\varepsilon(\tau)$ *is real and continuous with respect to* ε *satisfying*

$$\lim_{\varepsilon \downarrow 0} \lambda_c^\varepsilon(\tau) = \lambda_c^*(\tau).$$

Moreover, there exists a positive constant μ_2 *such that there are no real solutions of* $\mathcal{F}(\lambda; \varepsilon, \tau) = 0$ *in* $\mathbf{C}_{\mu_2}$ *for small* ε *except* $\lambda_c^\varepsilon(\tau)$ *and the zero eigenvalue coming from the translation invariance property.*

Proof. Recall that $\mathcal{F} = \mathcal{F}(\lambda; \varepsilon, \tau)$ is analytic with respect to λ in $\mathbf{C}_{\hat{\mu}}$ and continuous with respect to ε up to $\varepsilon = 0$. Since $\lambda = 0$ is always a solution of $\mathcal{F} = 0$ because of the translation invariance property, $\mathcal{F}$ can be decomposed as

$$\mathcal{F}(\lambda; \varepsilon, \tau) = \lambda H(\lambda; \varepsilon, \tau)$$

where H depends on $(\lambda; \varepsilon, \tau)$ in a similar way as $\mathcal{F}$ and H is real-valued for real λ. The critical eigenvalue $\lambda_c^*(\tau)$ of the SLEP equation (3.69) is a solution of $H(\lambda; 0, \tau) = 0$, namely,

$$H(\lambda_c^*(\tau); 0, \tau) = 0. \tag{3.71}$$

When $\lambda_c^*(\tau) \neq 0$, it follows from Lemmas 3.8 and 3.9 that the straight line S and the convex curve G intersect each other transversally at $\lambda = \lambda_c^*(\tau)$, which implies that

$$\frac{\partial \mathcal{F}}{\partial \lambda}(\lambda_c^*(\tau); 0, \tau) = \lambda_c^*(\tau) \frac{\partial H}{\partial \lambda}(\lambda_c^*(\tau); 0, \tau) \neq 0.$$

Therefore, it holds that

$$\frac{\partial H}{\partial \lambda}(\lambda_c^*(\tau); 0; \tau) \neq 0. \tag{3.72}$$

When $\lambda_c^*(\tau) = 0$ (the tangential case), it holds that

$$\frac{\partial \mathcal{F}}{\partial \lambda}(0; 0, \tau) = 0.$$

However, after some computation we have

$$\frac{\partial^2 \mathcal{F}}{\partial \lambda^2}(0; 0, \tau) = 2 \frac{\partial H}{\partial \lambda}(0; 0, \tau) = -2k^* \langle K_0^{*, \tau, c^*})^3 \delta_0, \delta_0 \rangle < 0.$$

Thus, (3.72) holds whether or not $\lambda_c^*(\tau) = 0$. Combining (3.71) with (3.72) and applying the Implicit Function Theorem to $H = 0$ (if necessary, we extend H to the negative ε continuously), we obtain the unique real critical eigenvalue $\lambda = \lambda_c^{\varepsilon}(\tau)$ of $\mathcal{F} = 0$ satisfying $\lim_{\varepsilon \downarrow 0} \lambda_c^{\varepsilon}(\tau) = \lambda_c^*(\tau)$.

Noting Lemma 3.9 and the uniqueness property of the Implicit Function Theorem, the latter part of Lemma 3.10 can be proved by contradiction without difficulty. Therefore we leave the details to the reader.

Lemma 3.10 is essential for judging the stability properties, since we have the following a priori bound for the remaining spectrum.

LEMMA 3.11. *Both the essential spectrum and the nonreal isolated eigenvalues of* $(LP)_{\varepsilon, \tau}$ *are uniformly bounded away from the imaginary axis for small* ε. *Namely, suppose that* λ *belongs to the above spectrum*; *then it holds that*

$$\lambda \notin \mathbf{C}_{\mu_3}, \tag{3.73}$$

for a positive constant μ_3 *independent of small* ε.

Proof. For the essential spectrum, this result is already proved in Proposition 3.1. For complex eigenvalues, we first show the following sublemma.

SUBLEMMA 3.1. *Suppose that* $\lambda \in \mathbf{C}_{\hat{\mu}}$ *is an isolated eigenvalue of* (3.68) (*for the definition of* $\hat{\mu}$, *see Remark* 3.5); *then for small* ε *it satisfies*

$$|\lambda| \leqq M_1$$

where M_1 *is a positive constant independent of* ε.

Proof. In view of Remark 3.6 and Lemma 3.6, we can easily prove by contradiction that there does not exist a divergent sequence of eigenvalues in $\mathbf{C}_{\hat{\mu}}$ of (3.68) as $\varepsilon \downarrow 0$. The details are left to the reader.

Using this sublemma, we will prove (3.73) for complex eigenvalues by contradiction. Suppose that there does not exist such a μ_3 for any small ε. It follows from Sublemma 3.1 that we can find a sequence of complex eigenvalues $\{\lambda_{\varepsilon_n}\}_{n\geqq 1}$ ($\operatorname{Im}\lambda_{\varepsilon_n} \neq 0$) of (3.68) with $\varepsilon_n \downarrow 0$ as $n \uparrow \infty$, which converges to λ_* satisfying the SLEP equation (3.69) with $\operatorname{Re}\lambda_* \geqq 0$. By recalling Lemma 3.7, the only possibility is that λ_* is equal to one of the real eigenvalues of (3.69). However, in the view of Lemma 3.10 and its proof, we see that both eigenvalues $\lambda_c^*(\tau)$ and the zero of (3.69) are uniquely extended as *real* solutions of (3.68), which is a contradiction and completes the proof of Lemma 3.11.

Except in the tangential case, i.e., $\lambda_c^*(\tau) = 0$, we can conclude the following from Lemmas 3.10 and 3.11.

THEOREM 3.2. *Let $\mathcal{U}^\varepsilon$ be a traveling front solution of $(P)_{\varepsilon,\tau}$ with the limiting velocity $c^* = \lim_{\varepsilon \downarrow 0} c(\varepsilon)$ constructed in § 2. Then, for small ε, $\mathcal{U}^\varepsilon$ is*

(i) *Asymptotically orbital stable in $\mathbf{C}_{\text{unif}}(\mathbf{R})$-topology if*

$$-\tau < \frac{d}{d\lambda} G(\lambda; \tau, c^*)\Big|_{\lambda=0} = -k^*\langle (K_0^{*,\tau,c^*})^2 \delta_0, \delta_0\rangle, \tag{3.74}$$

or

(ii) *Unstable if*

$$-\tau > \frac{d}{d\lambda} G(\lambda; \tau, c^*)\Big|_{\lambda=0} = -k^*\langle (K_0^{*,\tau,c^*})^2 \delta_0, \delta_0\rangle. \tag{3.75}$$

Proof of Theorem 3.2. It is clear from Lemmas 3.9 and 3.10 that the inequality $\tau >$ (respectively, $<$) $-(d/d\lambda)G(\lambda; \tau, c^*)|_{\lambda=0}$ implies the unique existence of the negative (respectively, positive) eigenvalue $\lambda_c^\varepsilon(\tau)$ for small ε. The simplicity of this critical eigenvalue $\lambda_c^\varepsilon(\tau)$ and the zero eigenvalue associated with the translation invariance can be proved in an analogous way to that of Theorem 4.1 of Nishiura and Mimura [13]. Therefore we omit the details. Thus, noting Lemma 3.11, we obtain the conclusion of Theorem 3.2.

4. Relation between stability and the sign of the Jacobian of matching conditions. In the previous section, we have obtained a stability criterion in which the location of the unique real critical eigenvalue is determined via comparison of slopes at the origin of two curves of the SLEP equation. In this section we will show that the stability criterion in Theorem 3.2 corresponds exactly to the *sign* of the Jacobian $\partial(\Phi_0, \Psi_0)/\partial(c, \beta)$, where Φ_0 and Ψ_0 are the functions used in C^1-matching conditions for singular limit solutions in § 2. In other words, the intersecting manner of two curves $\beta = \beta_0(c)$ and $\beta = \beta_I(c)$, associated with $\Phi_0 = 0$ and $\Psi_0 = 0$, respectively, determines the stability properties of traveling front solutions.

This not only gives us a clear geometrical interpretation of the stability criterion, but also has a useful practical application. In fact, when we construct the singular limit solutions, we can also judge their stabilities simultaneously. To prove this relation, we will prepare three lemmas: the first one deals with the formula of $(d/dc)\beta_0(c)$; similarly, $(d/dc)\beta_I(c)$ is computed in the second one; and, finally, in the third lemma a relation between the coefficient of the SLEP equation and the depth of the jump $j(\beta)$ at $\beta = \beta^*$ (see (4.1)) is given.

We introduce the following notation for later use:

$$G_\pm(V) \equiv g(h_\pm(V), V), \tag{4.1}_a$$

$$(4.1)_b \qquad G_\beta(V) \equiv \begin{cases} G_+(V) & \text{for } \beta \leqq V < v_{\max}, \\ G_-(V) & \text{for } v_{\min} < V < \beta, \end{cases}$$

and

$$(4.1)_c \qquad j(\beta) \equiv G_+(\beta) - G_-(\beta).$$

LEMMA 4.1. *Let $\beta = \beta_0(c)$ be the relation for β and c (see Lemma 2.2) where the outer equation $(2.7)_\pm$ for v has a unique C^1-matching solution $V_0(x; c)$ (see (2.10)). Then it holds that*

$$(4.2) \qquad \frac{d}{dc}\beta_0(c) = -j(\beta_0(c))^{-1}\left\{c(\beta_0(c) - v_-) - \int_{-\infty}^{0} g(U_0, V_0)\, dx\right\}^2 \langle z_0, z_0^* \rangle$$

where $U_0 \equiv h_{\beta_0(c)}(V_0)$, and z_0 (respectively, z_0^*) is given by $(d/dx) \cdot V_0(x; c)/(d/dx)V_0(0; c)$ (respectively, $e^{-cx}z_0$). Moreover, if $(c, \beta_0(c))$ is an intersection point with $\beta = \beta_I(c)$, then z_0 (respectively, z_0^*) is the limiting z-component of the eigenfunction associated with the translation free zero eigenvalue of $(LP)_{\varepsilon,\tau}$ (respectively, the adjoint problem $(LP)^*_{\varepsilon,\tau}$) under the normalization $z_0(0) = 1$ (respectively, $z_0^*(0) = 1$).

Proof. For arbitrary $c \in \mathbf{R}$, the outer solution V_0 matched in C^1-sense at the origin exists uniquely for $\beta = \beta_0(c)$ and satisfies the following equation (see Lemma 2.2):

$$(4.3) \qquad \frac{d^2}{dx^2}V_0 - c\frac{d}{dx}V_0 + G_{\beta_0(c)}(V_0) = 0.$$

Although (4.3) should be written in a weak form due to the jump discontinuity of $G_{\beta_0(c)}(\cdot)$ at the origin, we write it in a classical form for notational simplicity. Note that the C^1-matched solution V_0 of (4.3) can be regarded as a function of the velocity c; $V_0 = V_0(x; c)$. We rewrite (4.3) as

$$(4.4) \qquad \frac{d^2}{dx^2}V_0 - c\frac{d}{dx}V_0 + j(\beta_0(c))H_{\beta_0(c)}(V_0) + \hat{G}_{\beta_0(c)}(V_0) = 0$$

where $H_{\beta_0(c)}(\cdot)$ denotes the Heaviside function with unit jump at $\beta_0(c)$, and $\hat{G}_{\beta_0(c)}(\cdot) \equiv G_{\beta_0(c)}(\cdot) - j(\beta_0(c))H_{\beta_0(c)}(\cdot)$, which is a continuous function of x. Note that $H_{\beta_0(c)}(V_0(x))$ is the Heaviside function with unit jump at $x = 0$ as a function of x. Differentiating (4.4) with respect to x, we obtain

$$(4.5) \qquad \frac{d^2}{dx^2}(V_0)_x - c\frac{d}{dx}(V_0)_x + j(\beta_0(c))\delta_0(x) + \frac{d}{dV}\hat{G}_{\beta_0(c)}(V_0)(V_0)_x = 0.$$

On the other hand, differentiating (4.4) with respect to c, we have

$$\begin{aligned} &\frac{d^2}{dx^2}\dot{V}_0 - c\frac{d}{dx}\dot{V}_0 - \frac{d}{dx}V_0 + \frac{dj}{d\beta}(\beta_0(c))H_{\beta_0(c)}(V_0)\dot{\beta}_0(c) \\ &\quad - j(\beta_0(c))(V_0)_x(0)^{-1}\delta_0(x)\dot{\beta}_0(c) + j(\beta_0(c))\delta_0(x)\dot{V}_0 \\ &\quad + \frac{d}{d\beta}\hat{G}_{\beta_0(c)}(V_0)\dot{\beta}_0(c) + \frac{d}{dV}\hat{G}_{\beta_0(c)}(V_0)\dot{V}^0 = 0, \end{aligned} \qquad (4.6)$$

where $\cdot$ means c-differentiation. In view of (4.5), it is easily seen that $(V_0)^*_x = e^{-cx}(V_0)_x$ satisfies

$$(4.7) \qquad \frac{d^2}{dx^2}(V_0)^*_x + c\frac{d}{dx}(V_0)^*_x + j(\beta_0(c))\delta_0(x) + \frac{d}{dV}\hat{G}_{\beta_0(c)}(V_0)^*_x = 0.$$

This is apparently the adjoint equation of (4.5). Multiplying $(V_0)_x^{\#}$ to (4.6) on both sides, after integration by parts we obtain

$$
\begin{aligned}
&-\langle (V_0)_x, (V_0)_x^{\#}\rangle + \left\{ \left\langle \frac{dj}{d\beta}(\beta_0(c)) H_{\beta(c)}(V_0), (V_0)_x^{\#} \right\rangle \right. \\
&\qquad -\langle j(\beta_0(c))(V_0)_x(0)^{-1}\delta_0(x), (V_0)_x^{\#}\rangle \\
&\qquad \left. + \left\langle \frac{d}{d\beta}\hat{G}_{\beta_0(c)}(V_0), (V_0)_x^{\#} \right\rangle \right\} \dot{\beta}_0(c) = 0.
\end{aligned} \tag{4.8}
$$

We will show the following equality:

$$
\left\langle \frac{dj}{d\beta}(\beta_0(c)) H_{\beta_0(c)}(V_0), (V_0)_x^{\#} \right\rangle + \left\langle \frac{d}{d\beta}\hat{G}_{\beta_0(c)}(V_0), (V_0)_x^{\#} \right\rangle = 0. \tag{4.9}
$$

Making the Newton quotient of $\hat{G}_\beta$, we see that

$$
\begin{aligned}
\left\langle \frac{\hat{G}_{\beta+\Delta\beta}(V_0) - \hat{G}_\beta(V_0)}{\Delta\beta}, (V_0)_x^{\#} \right\rangle &= \frac{1}{\Delta\beta}\int_{x_\beta}^{x_{\beta+\Delta\beta}} \{\hat{G}_{\beta+\Delta\beta}(V_0) - \hat{G}_\beta(V_0)\}\cdot (V_0)_x^{\#}\, dx \\
&\quad - \frac{1}{\Delta\beta}\{j(\beta+\Delta\beta) - j(\beta)\}\int_{x_{\beta+\Delta\beta}}^{\infty} (V_0)_x^{\#}\, dx
\end{aligned} \tag{4.10}
$$

where x_β denotes the value of x where V_0 becomes β. Here, for definiteness, we only consider the case of $\Delta\beta > 0$. Note that x_β is uniquely defined because of the monotonicity of $V_0(x)$. Since $|\hat{G}_{\beta+\Delta\beta}(V_0(x)) - \hat{G}_\beta(V_0(x))| \leqq \max_{\beta'\in[\beta,\beta+\Delta\beta]} |j(\beta') - j(\beta)| \leqq$ const. $\Delta\beta$ and $x_{\beta+\Delta\beta}$ tends to x_β as $\Delta\beta \to 0$, the first term of the right-hand side of (4.10) goes to zero when $\Delta\beta \to 0$. Therefore, it follows from (4.10) that

$$
\left\langle \frac{d}{d\beta}\hat{G}_\beta(V_0), (V_0)_x^{\#} \right\rangle = -\frac{dj}{d\beta}\int_{x_\beta}^{\infty} (V_0)_0^{\#}\, dx,
$$

which shows (4.9) at $\beta = \beta_0(c)$. Substituting (4.9) into (4.8) and noting that $(V_0)_x = (V_0)_x^{\#}(0)$, we have

$$
\dot{\beta}_0(c) = -\frac{\langle (V_0)_x, (V_0)_x^{\#}\rangle}{j(\beta_0(c))} = -j(\beta_0(c))^{-1}((V_0)_x(0))^2\langle z_0, z_0^{\#}\rangle \tag{4.11}
$$

where z_0 (respectively, $z_0^{\#}$) is defined by $(V_0)_x/(V_0)_x(0)$ (respectively, $(V_0)_x^{\#}/(V_0)_x^{\#}(0)$). Let us compute the value of $(V_0)_x(0)$. Integrating (4.3) from $-\infty$ to 0, we obtain

$$
(V_0)_x(0) = c\{V_0(0) - v_-\} - \int_{-\infty}^{0} G_{\beta_0(c)}(V_0(x))\, dx.
$$

Since $x = 0$ is the switching point from the left branch $h_-(v)$ to the right one $h_+(v)$, and $V_0(0) = \beta_0(c)$, this becomes

$$
(V_0)_x(0) = c\{\beta_0(c) - v_-\} - \int_{-\infty}^{0} g(U_0, V_0)\, dx. \tag{4.12}
$$

Substituting this into (4.11), we obtain (4.2).

The last claim of Lemma 4.1 is clear from the fact that the x-derivative of the traveling front solution becomes an eigenfunction associated with the zero eigenvalue of $(LP)_{\varepsilon,\tau}$ for any small ε.

In view of Lemmas 2.3 and 2.4, the inner layer solutions were defined by solutions of the following stretched scalar equation:

$$W_{yy} - c\tau W_y + f(W, \beta) = 0, \qquad W(\pm\infty) = h_\pm(\beta),\ W(0) = \alpha \tag{4.13}$$

where $c = c_0(\beta)/\tau$ for any $\beta \in (v_-, v_+)$. By recalling Remark 2.1, the inner relation $c = c_0(\beta)/\tau$ can be solved with respect to β as $\beta = \beta_I(c)$. Here, of course, β_I depends also on τ; however, for simplicity, we do not write the τ-dependency explicitly. Therefore the solution of (4.13) can be regarded as a function of c for any fixed τ.

LEMMA 4.2. *Let $\beta = \beta_I(c)$ $(= \beta_I(c; \tau))$ be the inverse function of the inner C^1-matching condition $c = c_0(\beta)/\tau$ in Lemma* 2.4. *Then, it holds that*

$$\dot{\beta}_I(c) = \tau \Big/ \frac{d}{d\beta} c_0(\beta_I(c)). \tag{4.14}$$

Proof. Differentiating (4.13) with respect to c, we have

$$\dot{W}_{yy} - c\tau\dot{W}_y - \tau W_y + f_u(W, \beta_I(c))\dot{W} + f_v(W, \beta_I(c))\dot{\beta}_I(c) = 0 \tag{4.15}$$

where $\cdot$ means c-differentiation as before. On the other hand, differentiating (4.13) with respect to y, it is easily seen that W_y satisfies

$$\frac{d^2}{dy^2} W_y - c\tau \frac{d}{dy} W_y + f_u(W, \beta_I(c)) W_y = 0. \tag{4.16}$$

Therefore, $W_y^\# \equiv e^{-c\tau y} W_y$ satisfies

$$\frac{d^2}{dy^2} W_y^\# + c\tau \frac{d}{dy} W_y^\# + f_u(W, \beta_I(c)) W_y^\# = 0. \tag{4.17}$$

Note that both W_y and $W_y^\#$ decay exponentially as $|y| \to +\infty$ (see Lemma 2.3). Taking the inner product with $W_y^\#$ on both sides of (4.15) and using (4.17), after integration by parts we obtain

$$-\tau\langle W_y, W_y^\#\rangle + \dot{\beta}_I(c)\langle f_v(W, \beta_I(c)), W_y^\#\rangle = 0. \tag{4.18}$$

When we recall the relation (3.38), it follows from (4.18) that

$$\dot{\beta}_I(c) = \tau \Big/ \frac{d}{d\beta} c_0(\beta_I(c)),$$

which proves (4.14).

Now let (c^*, β^*) be an arbitrary transversal intersection point between $\beta = \beta_0(c)$ and $\beta = \beta_I(c)$, and $\mathcal{U}^\varepsilon = (u^\varepsilon, v^\varepsilon)$ be the corresponding traveling front solution. As we have remarked before, the spatial derivative $\mathcal{U}_x^\varepsilon = (u_x^\varepsilon, v_x^\varepsilon)$ satisfies $(LP)_{\varepsilon,\nu}$ with $\lambda = 0$ for any small ε. Therefore the SLEP differential equation (3.60) must also have zero eigenvalue and the z-component of the associated eigenfunction, under the normalization $\langle z_0, \delta_0\rangle = 1$, is given by

$$z_0 = \frac{k^*}{\hat{\zeta}_0^*} K_0^{*,\tau,c^*}\delta_0. \tag{4.19}$$

Recall that c^* is the limiting velocity of $\mathcal{U}^\varepsilon$ as $\varepsilon \downarrow 0$. Also noting that $z_0^\#$ $(= e^{-c^* x} z_0)$ is a kernel function of the adjoint problem (3.61), we see that it is represented by

$$z_0^\# = \frac{k^*}{\hat{\zeta}_0^*} K_0^{*,\tau,-c^*}\delta_0. \tag{4.20}$$

On the other hand, recalling that v^ε remains as a $C^1(\mathbf{R})$-function up to $\varepsilon = 0$, we see that v_x^ε converges to $(V_0)_x$. In view of (4.5) and Lemma 3.6, $(V_0)_x$ is represented by

$$(V_0)_x = j(\beta^*) K_0^{*,\tau,c^*} \delta_0. \tag{4.21}$$

Here we use the fact that $d\hat{G}_{\beta^*}/dV(V_0) = \det^*/f_u^*$ and $\beta^* = \lim_{\varepsilon \downarrow 0} v^\varepsilon(0)$. It is clear that z_0 $(=\lim_{\varepsilon \downarrow 0}(v_x^\varepsilon / v_x^\varepsilon(0)))$ and $(V_0)_x / \langle (V_0)_x, \delta_0 \rangle$ must coincide with each other. Thus in view of (4.19) and (4.21), we see that the following relation holds among the numbers k^*, $\hat{\zeta}_0^*$, and $j(\beta^*)$.

LEMMA 4.3. *Let (c^*, β^*) be an arbitrary intersection point of $\beta = \beta_0(c)$ and $\beta = \beta_I(c)$. Then it holds that*

$$\frac{j(\beta^*)}{(V_0)_x(0)} = \frac{k^*}{\hat{\zeta}_0^*}. \tag{4.22}$$

As has been seen in § 2, the singular limit traveling front solutions can be constructed by finding the intersection points between $\beta = \beta_0(c)$ and $\beta = \beta_I(c)$. We will show that the manner of intersection of these two curves is exactly equivalent to the sign of the real critical eigenvalue in Theorem 3.1. In other words, the inequalities $\dot{\beta}_I(c) \lesseqgtr \dot{\beta}_0(c)$ at $c = c^*$ are equivalent to those in Theorem 3.1. Our goal is the following.

THEOREM 4.1 (stability and matching conditions). *For a given $\tau > 0$, let (c^*, β^*) be an arbitrary intersection point of outer and inner relation curves $\beta = \beta_0(c)$ and $\beta = \beta_I(c)$ in § 2. Then, the inequalities*

$$\dot{\beta}_I(c^*) \gtreqless \dot{\beta}_0(c^*) \qquad \left(\cdot = \frac{d}{dc}\right) \tag{4.23}$$

are equivalent to

$$-\tau \gtreqless \frac{d}{d\lambda} G(\lambda; \tau, c^*)\Big|_{\lambda = 0} = -k^* \langle (K_0^{*,\tau,c^*})^2 \delta_0, \delta_0 \rangle, \tag{4.24}$$

which determine the stability properties of the traveling front solution. Moreover, the inequalities (4.23) (*or* (4.24)) *are equivalent to the sign of the Jacobian of C^1-matching conditions* $(\Phi_0(c, \beta; \tau), \Psi_0(c, \beta)) = (0, 0)$ (*see* § 2), *namely,*

$$\frac{\partial(\Phi_0, \psi_0)}{\partial(c, \beta)} \gtreqless 0 \quad \text{at } (c^*, \beta^*). \tag{4.25}$$

Proof. Using Lemmas 4.1 and 4.2, the inequalities (4.23) can be rewritten as

$$-\tau \gtreqless \frac{dc_0(\beta^*)}{d\beta} j(\beta^*)^{-1} \left\{ c^*(\beta^* - v_-) - \int_{-\infty}^{0} g(U_0, V_0)\, dx \right\}^2 \langle z_0, z_0^\# \rangle. \tag{4.26}$$

Recalling the formula (3.12) for $\hat{\zeta}_0^*$ and using Lemma 4.3 and (4.12), (4.26) becomes

$$-\tau \gtreqless -\frac{(\hat{\zeta}_0^*)^2}{k^*} \langle z_0, z_0^\# \rangle. \tag{4.27}$$

Substituting the expressions (4.19) and (4.20) into (4.27), we have

$$-\tau \gtreqless -k^* \langle K_0^{*,\tau,c^*} \delta_0, K_0^{*,\tau,-c^*} \delta_0 \rangle = -k^* \langle (K_0^{*,\tau,c^*}, \delta_0)^2, \delta_0 \rangle, \tag{4.28}$$

which is clearly equivalent to (4.24). Finally, we will prove the equivalence to (4.25). Rewriting the Jacobian as

$$\frac{\partial(\Phi_0, \Psi_0)}{\partial(c, \beta)} = \frac{\partial \Phi_0}{\partial \beta} \frac{\partial \Psi_0}{\partial \beta} \left\{ \frac{\partial \Phi_0 / \partial c}{\partial \Phi_0 / \partial \beta} - \frac{\partial \Psi_0 / \partial c}{\partial \Psi_0 / \partial \beta} \right\},$$

we see from the formula of derivative of the implicit function that, at the intersection point (c^*, β^*) of $\beta = \beta_0(c)$ and $\beta = \beta_I(c)$, it holds that

$$\left.\frac{\partial(\Phi_0, \Psi_0)}{\partial(c, \beta)}\right|_{(c,\beta)=(c^*,\beta^*)} = \frac{\partial \Phi_0}{\partial \beta}\frac{\partial \Psi_0}{\partial \beta}(\dot{\beta}_I(c^*) - \dot{\beta}_0(c^*)). \tag{4.29}$$

This combined with (2.9) and $(2.15)_b$ proves the last claim of Theorem 4.1.

Example 3.1. Suppose that $\beta = \beta_0(c)$ and $\beta = \beta_I(c)$ intersect each other as in Fig. 4(b). The traveling front solutions corresponding to P and Q (respectively, R) are stable (respectively, unstable).

Appendix. *Proof of Lemma* 3.7. The basic idea of the proof is essentially contained in the proof of Theorem 3.1 of Nishiura and Mimura [15]. However, it should be noted that $\hat{\zeta}_0^*$ and $G(\lambda; \tau, c^*)$ (see (3.12) and (3.70)) depend on the parameter τ, which is a different point from the case treated in [15].

For a given $\tau = \tau_0$, we must show the nonexistence of the complex eigenvalues in $\mathbf{C}_{\mu_1}$ for the SLEP equation (see (3.96) and (3.70)):

$$\mathcal{F}(\lambda; 0, \tau_0, c^*) = \hat{\zeta}_0^* - \tau_0\lambda - G(\lambda; \tau_0, c^*) = 0. \tag{1}$$

The strategy is that we consider the following modified equation of (1):

$$\tilde{\mathcal{F}}(\lambda; \tau) = \hat{\zeta}_0^* - \tau\lambda - G(\lambda; \tau_0, c^*) = 0, \tag{2}$$

namely, the coefficient of the second term of (1) is not fixed to be τ_0 and all the remaining terms are exactly the same as before. We study the behavior of the solutions of (2) when τ varies in $\mathbf{R}_+$, and prove the lemma by contradiction. For this purpose, we prepare four lemmas, the proofs of which will be given in the last part. First it follows from Lemma 3.8 that $\lambda = 0$ is always a solution of (2) and there exists a unique value $\tau = \tau_c$ (which depends on τ_0) such that the straight line $\hat{\zeta}_0^* - \tau_c\lambda$ is tangent to the convex curve $G(\lambda; \tau_0, c^*)$ at $\lambda = 0$. The first lemma describes the behavior of solutions of (2) near $(\tau, \lambda) = (\tau_c, 0)$ as follows.

LEMMA A1. *Equation* (2) *has exactly two solutions near* $(\tau, \lambda) = (\tau_c, 0)$. *One is the zero solution* $(\lambda \equiv 0)$, *which is independent of* τ, *and the other is the real solution* $\lambda \equiv \hat{\lambda}(\tau)$ *with* $\hat{\lambda}(\tau_c) = 0$, *which behaves, near* $\tau = \tau_c$, *as does*

$$\hat{\lambda}(\tau) \simeq -C_T(\tau - \tau_c) \tag{3}$$

where C_T *is a positive constant given by*

$$C_T \equiv \frac{1}{k^*\langle (K_0^{*,\tau_0,c^*})^3 \delta_0, \delta_0\rangle}.$$

Besides the zero solution, $\hat{\lambda}(\tau)$ *is a unique zero of* (2) *in an appropriate complex neighborhood of* $\lambda = 0$.

Remark A1. When $\tau \neq \tau_c$, it holds that $\partial\tilde{\mathcal{F}}/\partial\lambda(\hat{\lambda}(\tau); \tau) \neq 0$ at the unique nonzero real solution $\lambda = \hat{\lambda}(\tau)$ of (2).

The next lemma shows the uniqueness of solutions of (2) on the imaginary axis.

LEMMA A2. *Suppose there exists a solution of* (2) *that crosses the imaginary axis when* τ *varies. Then it must be a real one and must coincide with the solution* $\hat{\lambda}(\tau)$ *in Lemma* A1. *Therefore, there are no complex solutions with* Im-*part* $\neq 0$ *that cross the imaginary axis.*

The following lemma shows that a complex solution of (2) can always be extended uniquely as a function of τ.

LEMMA A3. *Let* (λ_1, τ_1) *be a solution of* (2) *in* $\mathbf{C}_{\hat{\mu}} \times \mathbf{R}_+$. *If* $\operatorname{Im} \lambda_1 \neq 0$, *then it holds that*

$$\frac{\partial \tilde{\mathcal{F}}}{\partial \lambda}(\lambda_1; \tau) \neq 0.$$

Finally, the next lemma shows the nonexistence of solutions of (2) for large τ.

LEMMA A4. *There exist positive constants* τ_s *and* μ_s *such that there are no solutions of* (2) *in* $\mathbf{C}_{\mu_s}$ *for* $\tau \geqq \tau_s$ *except the simple zero solution.*

Now we are ready to prove Lemma 3.7 by contradiction. Suppose that Lemma 3.7 is not true; then we can find a nonreal solution $\lambda_0(\tau_0)$ of (1) with $\operatorname{Re}(\lambda_0(\tau_0)) \geqq 0$. Here we use the fact that $\lambda = 0$ is not an accumulation point of solutions of (1), which can be easily checked by using the properties of G (see Lemma 3.8) and the proof of Lemma A1. Note that τ_0 must be strictly smaller than τ_s in Lemma A4. Regarding $\lambda_0(\tau_0)$ as a special solution of (2) for $\tau = \tau_0$, we trace its behavior when τ varies. We denote by $\lambda_0(\tau)$ the solution of (2), which is a continuation of $\lambda_0(\tau_0)$ as τ varies. In view of Lemma A3, $\lambda_0(\tau_0)$ is uniquely continued as far as it remains a nonreal solution. When τ increases, we see from Lemma A4 that $\lambda_0(\tau)$ must cross the imaginary axis before τ reaches τ_s. However, owing to Lemma A2, $\lambda_0(\tau)$ cannot cross the imaginary axis. Therefore $\lambda_0(\tau)$ must fall into the real solution of (2) before it reaches the imaginary axis. But we see from Lemma A1 and Remark A1 that this is not possible, which is a contradiction and completes the proof.

Proof of Lemma A1. It is clear from Lemma 3.8(i) that $\lambda = 0$ is always a solution of (2) independent of τ. Therefore, since $\tilde{\mathcal{F}}$ is analytic with respect to λ, $\tilde{\mathcal{F}}$ can be rewritten locally near $\lambda = 0$ in the following form:

$$\tilde{\mathcal{F}}(\lambda; \tau) = \lambda \tilde{H}(\lambda; \tau) \tag{4}$$

where $\tilde{H}$ is a smooth function with respect to all variables. After some computation, we obtain

$$\frac{\partial^2 \tilde{\mathcal{F}}}{\partial \lambda^2}(0; \tau_c) = -2k^* \langle (K_0^{*, \tau_0, c^*})^3 \delta_0, \delta_0 \rangle < 0. \tag{5}$$

Since $\partial^2 \tilde{\mathcal{F}} / \partial \lambda^2 (0; \tau_c) = 2 \partial \tilde{H} / \partial \lambda (0, \tau_c)$, it follows from (5) that

$$\frac{\partial \tilde{H}}{\partial \lambda}(0; \tau_c) = -k^* \langle (K_0^{*, \tau_0, c^*})^3 \delta_0, \delta_0 \rangle < 0. \tag{6}$$

This implies via the Implicit Function Theorem that $\tilde{H} = 0$ has a unique solution $\lambda = \hat{\lambda}(\tau)$ in an appropriate complex neighborhood of $\lambda = 0$ with $\hat{\lambda}(\tau_c) = 0$. It also holds that

$$\left. \frac{d\hat{\lambda}(\tau)}{d\tau} \right|_{\tau = \tau_c} = -\frac{\partial \tilde{H} / \partial \tau (0; \tau_c)}{\partial \tilde{H} / \partial \lambda (0; \tau_c)} = -\frac{1}{k^* \langle (K_0^{*, \tau_0, c^*})^3 \delta_0, \delta_0 \rangle}, \tag{7}$$

which completes the proof of Lemma A1.

Proof of Lemma A2. We will show that the origin $(\lambda = 0)$ is the unique solution on the imaginary axis of (2). The real and imaginary parts of (2) are given by

$$\hat{\zeta}_0^* - \tau \lambda_R - A(\lambda_R, (\lambda_I)^2) = 0, \tag{8}_a$$

$$B(\lambda_R, (\lambda_I)^2) - \tau = 0 \tag{8}_b$$

where $\lambda = \lambda_R + i\lambda_I$, and A and B are smooth functions of λ_R $(> -\hat{\mu})$ and $(\lambda_I)^2 \geqq 0$ defined by

$$(9)_a \qquad A(\lambda_R, (\lambda_I)^2) = k^*\langle \hat{I}K_{\lambda_R}\delta_0, \delta_0\rangle,$$

$$(9)_b \qquad B(\lambda_R, (\lambda_I)^2) = k^*\langle \hat{I}K_{\lambda_R}^2\delta_0, \delta_0\rangle.$$

Here $K_{\lambda_R}: (H_\rho^1)^* \to H_\rho^1$ is the inverse operator of

$$T_{\lambda_R} \equiv -\frac{d^2}{dx^2} - c^*\frac{d}{dx} - \frac{\det^*}{f_u^*} + \lambda_R \quad \text{and} \quad \hat{I} = \{I + (\lambda_I)^2 K_{\lambda_R}^2\}^{-1}: H_\rho^1 \to H_\rho^1.$$

The following results are useful, the proofs of which are left to the reader (see also Lemma 3.2 of Nishiura and Mimura [15]).

SUBLEMMA A1.

(i) $\dfrac{\partial A}{\partial(\lambda_I)^2} < 0$ *and* $\dfrac{\partial B}{\partial(\lambda_I)^2} < 0,$

(ii) $\lim\limits_{|\lambda_I| \uparrow \infty} A = 0$ *and* $\lim\limits_{|\lambda| \uparrow \infty} B = 0,$

(iii) $\dfrac{\partial B}{\partial \lambda_R} < 0$ *for* $\lambda_R > -\hat{\mu},$

(iv) $B(\lambda_R, 0) = -\dfrac{d}{d\lambda_R} A(\lambda_R, 0).$

Let us set $\lambda_R = 0$ in (8) and $B_0 = B(0, 0)$. It follows from Sublemma A1(i) that $(8)_b$ can be solved uniquely with respect to $(\lambda_I)^2$ as a function of τ for $\tau \leqq B_0$. We denote it by $(\lambda_I)^2(\tau)$. Note that $(\lambda_I)^2(\tau)$ is a strictly decreasing function of τ with $(\lambda_I)^2(B_0) = 0$ and $\lim_{\tau \to 0} (\lambda_I)^2(\tau) = +\infty$. Substituting this into $(8)_a$, we have a scalar equation of τ $(\leqq B_0)$:

$$(10) \qquad \hat{\zeta}_0^* - A(0, (\lambda_I)^2(\tau)) = 0.$$

In view of Sublemma A1(i), we see that $A(0, (\lambda_I)^2(\tau))$ is a strictly monotone increasing function of τ. Therefore (10) has a unique solution, if it exists. On the other hand, we already know that $\lambda = 0$ is a special solution of (8) coming from the translation invariance, which shows that $\tau = B_0$ is a unique zero of (10). Thus, all the eigenvalues that cross the imaginary axis must go through the origin. Recalling the local uniqueness of the zero solution of (2) in some complex neighborhood (see Lemma A1), we see that $\hat{\lambda}(\tau)$ is the unique solution of (2) which crosses the imaginary axis as τ varies.

Proof of Lemma A3. This can be done in the spirit of the proof of Proposition 3.1 of Nishiura and Mimura [15]. So we leave the details to the reader.

Proof of Lemma A4. First note that, from Sublemma A1(i) and (iii), B is a strictly decreasing function λ_R and $(\lambda_I)^2$. Let $\tau_s \equiv B(-\hat{\mu}/2, 0)$. Then, in view of $(8)_b$, we see that there are no complex eigenvalues in the region $\mathrm{Re}\,\lambda \geqq -\mu_1/2$ for $\tau \geqq \tau_s$. Recalling Sublemma A1(iv) and $A(\lambda_R, 0) \equiv G(\lambda_R; \tau_0, c^*)$ (see (3.70)), we see that $-\tau_s$ is the slope of the convex curve $G(\lambda; \tau_0, c^*)$ at $\lambda = -\hat{\mu}/2$. A simple geometric consideration implies that there are no real eigenvalues λ satisfying $\lambda \geqq -\hat{\mu}/2$ for $\tau \geqq \tau_s$. Combining these results, we can conclude Lemma A4 with $-\mu_s = \max\{-\mu_1/2, -\delta_e\}$ for $\tau \geqq \tau_s$ (see Proposition 3.1 for the definition of δ_e).

REFERENCES

[1] E. A. CODDINGTON and N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[2] D. E. EDMUNDS AND W. D. EVANS, *Spectral Theory and Differential Operators*, Clarendon Press, Oxford, 1987.

[3] J. W. EVANS, *Nerve axon equations*, III: *stability of the nerve impulse*, Indiana Univ. Math. J., 22 (1972), pp. 577-593.

[4] ———, *Nerve axon equations*, IV: *the stable and the unstable impulse*, Indiana Univ. Math. J., 24 (1975), pp. 1169-1190.

[5] P. C. FIFE, *Boundary and interior transition layer phenomena for pairs of second-order differential equations*, J. Math. Anal. Appl., 54 (1976), pp. 497-521.

[6] P. C. FIFE AND J. B. MCLEOD, *The approach of solutions of nonlinear diffusion equation to travelling wave solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 335-361.

[7] S. GOLDBERG, *Unbounded Linear Operators: Theory and Applications*, McGraw-Hill, New York, 1966.

[8] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Mathematics 80, Springer-Verlag, Berlin, New York, 1981.

[9] Y. HOSONO AND M. MIMURA, *Singular perturbation approach to traveling waves in competing and diffusing species models*, J. Math. Kyoto Univ., 22 (1982), pp. 435-461.

[10] H. IKEDA, M. MIMURA, AND Y. NISHIURA, *Global bifurcation phenomena of traveling wave solutions for some bistable reaction-diffusion systems*, Nonlinear Anal. Theory Methods Appl., 13 (1989), pp. 507-526.

[11] Y. NISHIURA, *Stability analysis of travelling wave solutions of reaction-diffusion systems—An application of the* SLEP *method*, manuscript.

[12] ———, *Singular limit approach to stability and bifurcation for bistable reaction diffusion systems*, Proc. Workshop on Nonlinear PDEs, March 1987, Provo, Utah, P. Bates and P. Fife, eds., Rocky Mountain J. Math., to appear.

[13] Y. NISHIURA AND H. FUJII, *Stability of singularly perturbed solutions to systems of reaction-diffusion equations*, SIAM J. Math. Anal., 18 (1987), pp. 1726-1770.

[14] ———, SLEP *method to the stability of singular perturbed solutions with multiple internal transition layers in reaction-diffusion systems*, Dynamics of Infinite Dimensional Systems, S.-N. Chow and J. K. Hale, eds., NATO ASI Series, F37, Springer-Verlag, Berlin, New York, 1987, pp. 221-230.

[15] Y. NISHIURA AND M. MIMURA, *Layer oscillations in reaction-diffusion systems*, SIAM J. Appl. Math., 49 (1989), pp. 481-514.

[16] J. RINZEL AND D. TERMAN, *Propagation phenomena in a bistable reaction-diffusion system*, SIAM J. Appl. Math., 42 (1982), pp. 1111-1137.

[17] C. JONES AND R. GARDNER, private communication.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 123-136, January 1990
© 1990 Society for Industrial and Applied Mathematics
007

TRAVELLING WAVE SOLUTIONS TO A SEMILINEAR DIFFUSION SYSTEM*

J. ESQUINAS† AND M. A. HERRERO†

Abstract. This paper considers the semilinear system

$$\text{(S)}\qquad \begin{aligned} &u_t - u_{xx} + v^p = 0,\\ &v_t - v_{xx} + u^q = 0,\\ &-\infty < x < +\infty, \qquad t > 0 \end{aligned}$$

with $p > 0$ and $q > 0$, and looks for nonnegative and nontrivial travelling wave solutions to (S): $u(x, t) = \varphi(ct - x)$, $v(x, t) = \psi(ct - x)$ possessing sharp fronts, i.e., such that $\varphi(\xi) = \psi(\xi) = 0$ for $\xi \leqq \xi_0$ and some finite ξ_0, which after a phase shift can always be assumed to be located at the origin. These solutions are called finite travelling waves (FTW). Here it is shown that if $pq < 1$, for any real c there exists an FTW that is unique up to phase translations and unbounded, whereas no FTW exists if $pq \geqq 1$. The asymptotic wave profiles near the front as well as far from it are also determined.

Key words. semilinear diffusion systems, travelling waves, fronts, asymptotic behaviour

AMS(MOS) subject classifications. 35K55, 35K57, 35R35

1. Introduction.

In this paper we will consider the system

$$\text{(1.1)}\qquad \begin{aligned} &u_t - u_{xx} + v^p = 0,\\ &v_t - v_{xx} + u^q = 0,\\ &-\infty < x < +\infty, \qquad t > 0 \end{aligned}$$

where p and q are positive real numbers. More precisely, we are interested in the existence of nonnegative *finite travelling waves* (FTW). By a *travelling wave* of (1.1) with speed c we mean a solution $(u(x, t), v(x, t))$ defined in

$$S = \{(x, t): -\infty < x < +\infty, t > 0\}$$

of the form

$$\text{(1.2)}\qquad u(x, t) = \varphi(ct - x), \qquad v(x, t) = \psi(ct - x)$$

where $\varphi(\xi)$ and $\psi(\xi)$ are nonnegative and different from zero, $\varphi, \psi \in \mathscr{C}^2(-\infty, +\infty)$, and $\varphi(\xi), \psi(\xi) \to 0$ as $\xi \to -\infty$ and are nondecreasing in ξ. Here φ and ψ are the respective wave profiles, and c may be any real number. If $\varphi(\xi) = \psi(\xi) = 0$ when $\xi \leqq \xi_0$ for some real ξ_0 we say that (u, v) is a *finite travelling wave.* In this case the line $x = ct - \xi_0$ is a front separating the region $P_+(u, v) = \{(x, t): u > 0, v > 0\}$ from the one where $u = v = 0$. Clearly, $P_+(u, v)$ expands in time when $c > 0$ and recedes if $c < 0$, to remain stationary for $c = 0$. In the context of scalar heat conduction problems the two first cases are referred to as the onset of heating and cooling waves, respectively.

There exists a wide literature on reaction-diffusion systems and their stationary states (cf., for instance, [A], [S], and references therein), and, in particular, on the existence of travelling waves to them (see [AW], [F], [FM], [BNS], [CL], [H], [T],

* Received by the editors January 13, 1988; accepted for publication (in revised form) February 14, 1989. This work was partly supported by Comisión Interministerial para la Ciencia y la Tecnologia grant PB86-0112-C0202.

† Departamento de Matemática Aplicada, Facultad de Matemáticas, Universidad Complutense, 28040 Madrid, Spain.

and the references therein). Here we will concern ourselves with one of the simplest cases of semilinear systems exhibiting a nontrivial coupling involving zeroth-order terms, but the techniques employed extend to a wide class of diffusion-absorbing problems. See in this context the remark at the end of § 4.

Let us proceed to describe our results. As the nonlinear terms v^p, u^q are monotone and unbounded for nonnegative values of u and v, no travelling wave can be expected that connects zero to a positive constant state. Thus our FTW will be unbounded, a fact already observed in nonlinear heat conduction problems in the scalar case (cf. [M], [HV]). We first obtain a necessary and sufficient condition for FTW's to exist in terms of p and q, namely, Theorem 1 below.

THEOREM 1. *There exist finite travelling wave solutions of* (1.1) *if and only if*

$$pq < 1. \tag{1.3}$$

Moreover, if (1.3) *holds, for any real c there exists an* FTW *moving with speed c, and the corresponding wave profiles* φ *and* ψ *are unique up to translations in space and time.*

We will assume henceforth that (1.3) holds, and proceed to derive the asymptotic wave profiles near the front (which we may assume to be located at $\xi = ct - x = 0$) as well as for large values of ξ. To this end, we recall some notation. We say that $f(\xi)$ and $g(\xi)$ are equivalent as $\xi \to \xi_0$ (finite or infinite), and write $f(\xi) \approx g(\xi)$ as $\xi \to \xi_0$, if $\lim_{\xi \to \xi_0} (f(\xi)/g(\xi)) = 1$. We then have Theorem 2.

THEOREM 2. *Assume that* $pq < 1$ *and for any real c, let* (φ, ψ) *be the* FTW *propagating with speed c obtained in Theorem* 1. *Then the following hold.*

(i) *For any real c, we have*

$$\varphi(\xi) \approx A\xi^{\alpha} \text{ and } \varphi(\xi) \approx B\xi^{\beta} \text{ as } \xi \to 0^+, \text{ where} \tag{1.4}$$

$$\alpha = \frac{2(1+p)}{1-pq}, \qquad \beta = \frac{2(1+q)}{1-pq},$$

$$A^{1-pq} = ((\beta(\beta-1))^p \alpha(\alpha-1))^{-1}, \qquad B = A^q(\beta(\beta-1))^{-1}.$$

(ii) *If* $c < 0$, *then*

$$\varphi(\xi) \approx C\xi^{\gamma} \text{ and } \psi(\xi) \approx D\xi^{\delta} \text{ as } \xi \to \infty, \text{ where} \tag{1.5}$$

$$\gamma = \frac{1+p}{1-pq}, \qquad \delta = \frac{1+q}{1-pq},$$

$$C^{1-pq} = ((-c)^{1+p}\delta^p\gamma)^{-1}, \qquad D = C^q((-c)\delta)^{-1}.$$

(iii) *If* $c > 0$, *the asymptotics for* $\xi \to \infty$ *depend on the values of p and q as follows*:

If $p < 1$, $q < 1$ *then* $\varphi(\xi) \approx M_1 e^{c\xi}$, $\psi(\xi) \approx N_1 e^{c\xi}$, *where* (1.6a)

$$M_1^{1-pq} = \frac{1}{(c^2(1-q))^p c^2(1-p)}, \qquad N_1 = \frac{M_1^q}{c^2(1-q)};$$

If $p < 1$, $q = 1$ *then* $\varphi(\xi) \approx M_2 e^{c\xi}$, $\psi(\xi) \approx N_2 \xi e^{c\xi}$, *where* (1.6b)

$$M_2^{1-p} = \frac{1}{c^{p+1}} \int_0^{\infty} s^p e^{c(p-1)s}\, ds, \qquad N_2 = \frac{M_2}{c};$$

If $p < 1$, $q > 1$ *then* $\varphi(\xi) \approx M_3 e^{c\xi}$, $\psi(\xi) \approx N_3 e^{cq\xi}$, *where* (1.6c)

$$M_3^{1-pq} = \frac{1}{(c^2(q-1)q)^p c^2(1-pq)}, \qquad N_3 = \frac{M_3^q}{c^2(q-1)q}.$$

The cases $p>1$, $q<1$ and $p>1$, $q=1$ are obtained by changing the coefficients in (1.6b, c) *in an obvious way.*

Let us remark briefly on the estimates above. First, when we look for solutions of the form (1.2), (1.1) is reduced to an ordinary differential equation (ODE) system in the variable $\xi = ct - x$, namely,

$$\varphi'' = c\varphi' + \psi^p, \qquad \psi'' = c\psi' + \varphi^q. \tag{1.7}$$

When $c=0$ (i.e., in the case of stationary waves), (1.7) has an explicit solution given by

$$\varphi(\xi) = A\xi^\alpha, \quad \psi(\xi) = B\xi^\beta, \quad \alpha, \beta, A, B, \text{ as in (1.4).} \tag{1.8}$$

Therefore (1.4) shows that, regardless of the wave-speed c, the first-order asymptotics of FTW's near $\xi=0$ is precisely that of the stationary solutions (1.8). As for the behaviour when $\xi\to\infty$, consider beginning with the case $c<0$ and try formally the asymptotic expansion

$$\varphi(\xi) = C\xi^\gamma + \cdots, \quad \psi(\xi) = D\xi^\delta + \cdots, \quad \xi \gg 0.$$

Substituting this into (1.7) yields

$$\begin{aligned} C\gamma(\gamma-1)\xi^{\gamma-2} &= cC\gamma\xi^{\gamma-1} + D^p\xi^{\delta p} + \cdots, \\ D\delta(\delta-1)\xi^{\delta-2} &= cD\delta\xi^{\delta-1} + C^q\xi^{\gamma q} + \cdots \end{aligned} \tag{1.9}$$

so that, neglecting the terms on the left of (1.9) and matching those on the right, we obtain the values listed in (1.5). We call the wave profiles thus determined to be of an *absorptive nature*, since constants C, D, δ, γ in (1.5) are those corresponding to the explicit FTW solution of the simplified absorbing system obtained from (1.1) by dropping the terms u_{xx}, v_{xx} there.

When $c>0$ we expect the leading behaviour at infinity to be influenced indeed by the heatlike part of (1.1). For instance, if $p<1$ and $q>1$, trying in (1.7) $\varphi(\xi)\approx P e^{\mu\xi}$ and $\psi(\xi)\approx Q e^{\nu\xi}$ leads to

$$P\mu^2 e^{\mu\xi} = cP\mu\, e^{\mu\xi} + Q^p\, e^{p\nu\xi} + \cdots, \tag{1.10a}$$

$$Q\nu^2 e^{\nu\xi} = cQ\nu\, e^{\nu\xi} + P^q\, e^{\mu q\xi} + \cdots. \tag{1.10b}$$

Considering then the three terms in (1.10b) as of the same order yields $\nu = \mu q$, whereas matching the first two in (1.10a) gives us $\mu = c$. We then obtain $\nu = cq$, $\mu = c$ and the third term in (1.10a) is of lower order as $\xi\to\infty$. These are the wave behaviours stated in (1.6c), except for the coefficients P, Q, which remain to be obtained. A similar analysis can be performed for the remaining cases, always yielding the exponential-like estimates for φ and ψ listed in (1.6). Because of the influence of the parabolic part of (1.1), we call these behaviours of *diffusive* type, in contrast wih those obtained in (1.4), (1.5).

As to the precedents of this paper, we should first mention the work [RK], where the authors considered the general scalar equation

$$u_t = a(u^m)_{xx} - bu^p + k(u^n)_x, \qquad a, m, n, b, p, k > 0 \tag{1.11}$$

and used formal perturbation theory (as in the remarks just made above), to describe all the possible asymptotic behaviours of waves occurring in (1.11) in terms of the different parameters therein. See also [PP], where the onset of FTW for another kind of degenerate diffusion equation is analyzed in a similar way. Later, in [HV], a rigorous justification of the results conjectured in [RK] was provided for the case $k=0$; negative

values of m and p were also allowed. The methods employed involve phase space and comparison arguments. We have resorted to different techniques here, since the natural phase space associated with (1.7) is four-dimensional, and therefore rather unwieldy. Basically, existence is obtained by a fixed-point argument in whose formulation the asymptotic behaviours predicted by perturbation theory play a crucial role (cf. § 3). As to uniqueness, it follows from a general ODE argument as explained in § 2 below.

From the results in [RK] and [HV] it follows, in particular, that for the equation

$$u_t = a(u^m)_{xx} - bu^p, \quad a, b > 0, \quad m, p > -1 \tag{1.12}$$

only three asymptotic behaviours of FTW's are possible, there named as *absorptive*, *diffusive*, and *stationary*. These are illustrated by explicit solutions easily obtained by dropping, respectively, the second, third, and first terms in (1.12) (cf. [HV, § 2]). In particular, for the semilinear equation

$$u_t - u_{xx} + u^p = 0, \tag{1.13}$$

FTW's $u(x, t) = \varphi(ct - x)$ exist for positive values of p if and only if $p < 1$. Their behaviour is stationary ($\varphi(\xi) \approx A_1 \xi^{(2/1-p)}$ near the front $\xi \equiv ct - x = 0$, and absorptive ($\varphi(\xi) \approx A_2 \xi^{(1/1-p)}$ or diffusive ($\varphi(\xi) \approx A_3 e^{c\xi}$) as $\xi \to \infty$, according, respectively, to whether $c < 0$ or $c > 0$. Here A_1 and A_2, are positive constants, A_2 depends on c but A_1 does not, and their explicit values are determined, for instance, by trying a formal perturbation expansion. On the other hand,

$$A_3 = \left(\frac{1}{c^2(1-p)}\right)^{(1/1-p)},$$

as can be seen by repeating the arguments in § 4 below. The reader will notice the analogy between these results and the discussion following Theorems 1 and 2 above, although several diffusive behaviours are possible for (1.1) depending on the values of p and q.

Higher-order wave asymptotics can also be obtained from our techniques. As an example, we derive a two-term expansion for FTW's near the fronts.

THEOREM 3. *Let $pq < 1$, and for any real c let $u_c(x, t) = \varphi(ct - x)$, $v_c(x, t) = \psi(ct - x)$ be the solution obtained in Theorems 1 and 2. Then, for $ct - x$ positive and close enough to zero,*

$$u_c(x, t) \approx A(ct - x)^{\alpha} + A_1(ct - x)^{\alpha+1} + \cdots, \tag{1.14}$$

$$v_c(x, t) \approx B(ct - x)^{\beta} + B_1(ct - x)^{\beta+1} + \cdots \tag{1.15}$$

where A, B, α, β are as in Theorem 2 and

$$A_1 = \frac{cA(p(\alpha - 1) + \beta + 1)}{(\alpha + 1)(\beta + 1) - pq(\beta - 1)(\alpha - 1)}, \tag{1.16}$$

$$B_1 = \frac{cB\beta + qA^{q-1}A_1}{\beta(\beta + 1)}. \tag{1.17}$$

Note that the wave-speed c, which is not reflected in the first-order approximation near the front, does appear when second-order terms are considered.

We conclude this Introduction with the plan of the paper. Some preliminaries, as well as uniqueness and nonexistence results, are gathered in § 2. Section 3 is devoted to existence and asymptotics of cooling waves ($c < 0$) under the necessary condition $pq < 1$. Section 4 covers the case $c > 0$.

2. Preliminaries. Nonexistence when $pq \geqq 1$. Uniqueness. A finite travelling wave to (1.1) can be described as a nontrivial, nonnegative solution to the ODE system

$$\begin{aligned} &\varphi, \psi \in C^2[0, \infty), \\ &\varphi''(\xi) = c\varphi'(\xi) + \psi(\xi)^p \quad \text{for } \xi > 0, \\ &\psi''(\xi) = c\psi'(\xi) + \varphi(\xi)^q \quad \text{for } \xi > 0 \end{aligned} \tag{2.1}$$

together with conditions

$$\varphi(\xi) = \varphi'(\xi) = \psi(\xi) = \psi'(\xi) = 0 \quad \text{for } \xi \leqq 0. \tag{2.2}$$

If such a solution exists, the following representation formulas hold:

$$\varphi(\xi) = \int_0^\xi \left(\frac{e^{c(\xi-s)}-1}{c}\right) \psi(s)^p \, ds, \tag{2.3a}$$

$$\psi(\xi) = \int_0^\xi \left(\frac{e^{c(\xi-s)}-1}{c}\right) \varphi(s)^q \, ds, \tag{2.3b}$$

$$\varphi'(\xi) = \int_0^\xi e^{c(\xi-s)} \varphi(s)^p \, ds, \tag{2.3c}$$

$$\psi'(\xi) = \int_0^\xi e^{c(\xi-s)} \varphi(s)^q \, ds. \tag{2.3d}$$

Indeed, (2.3c, d) can be obtained just by differentiating in the first two equations above. Now, substituting (2.3b) into (2.3a), we get

$$\varphi(\xi) = \int_0^\xi \left(\frac{e^{c(\xi-t)}-1}{c}\right) \left[\int_0^t \left(\frac{e^{c(t-s)}-1}{c}\right) \varphi(s)^q \, ds\right]^p dt. \tag{2.4}$$

Therefore, finding FTW's amounts to obtaining nontrivial solutions to the integral equation (2.4).

We now state a nonexistence result.

LEMMA 2.1. *There are no* FTW's *when* $pq \geqq 1$.

Proof. Let us show that $\varphi(\xi) = \psi(\xi) \equiv 0$ if $pq \geqq 1$. If $p \geqq 1$, and $q \geqq 1$, the conclusion follows at once from standard uniqueness results. As to the general case, we will argue by contradiction. By (2.3c, d) both φ and ψ are monotone nondecreasing, and once they are different from zero they always stay positive. Without loss of generality, we may then assume $\varphi(\xi) > 0$, $\psi(\xi) > 0$ for $\xi > 0$. Then from (2.4) we deduce

$$\varphi(\xi) \leqq \varphi(\xi)^{pq} F(\xi) \tag{2.5}$$

where

$$F(\xi) = \int_0^\xi \left(\frac{e^{c(\xi-t)}-1}{c}\right) \left[\int_0^t \left(\frac{e^{c(t-s)}-1}{c}\right) ds\right]^p dt$$

so that, in particular, $F(\xi) \to 0$ as $\xi \to 0$. Now the contradiction follows, since by (2.5), $\varphi(\xi)^{1-pq}$ should go to zero as $\xi \to 0$, which is impossible. □

Uniqueness up to phase shift is contained in the following lemma.

LEMMA 2.2. *Suppose that* (φ_1, ψ_1) *and* (φ_2, ψ_2) *are nontrivial solutions of* (2.1), (2.2). *Then for some real* η *we have* $\varphi_2(\xi) = \varphi_1(\xi - \eta)$ *and* $\psi_2(\xi) = \psi_1(\xi - \eta)$.

Proof. The argument involves two steps.

(i) Any pair of different nontrivial solutions (φ_1, ψ_1) and (φ_2, ψ_2) is ordered, i.e., $\varphi_1(\xi) < \varphi_2(\xi)$ and $\psi_1(\xi) < \psi_2(\xi)$ or $\varphi_1(\xi) > \varphi_2(\xi)$ and $\psi_1(\xi) > \psi_2(\xi)$ for $\xi > 0$.

Actually, it suffices to show the result for φ_1 and φ_2, since the corresponding statement for ψ_1, ψ_2 then follows from (2.3b). To begin with, if supp $\varphi_1 \subset$ supp φ_2 and supp $\varphi_1 \neq$ supp φ_2, (2.3a, b) yield $\varphi_1(\xi) > \varphi_2(\xi)$ for all $\xi > 0$. Therefore we have only to deal with the case

$$\varphi_1(0) = \varphi_2(0), \quad \varphi_1(\xi) > 0, \quad \varphi_2(\xi) > 0 \quad \text{for all } \xi > 0.$$

We then argue by contradiction as follows. Assume that there exists $\xi_0 > 0$ such that $\varphi_1(\xi_0) = \varphi_2(\xi_0)$. Then there exists $\eta > 0$ such that, for any $\xi \leqq \xi_0$,

$$\varphi_{1,\eta}(\xi) \leqq \varphi_2(\xi)$$

where

$$\varphi_{1,\eta}(\xi) = \begin{cases} 0 & \text{if } \xi \leqq \eta, \\ \varphi_1(\xi - \eta) & \text{if } \xi \geqq \eta. \end{cases}$$

Note that for arbitrary $\eta \geqq 0$, $(\varphi_\eta(\xi), \psi_\eta(\xi))$ is a solution of (2.1) provided $(\varphi(\xi), \psi(\xi))$ is also. Now we set $\mu_0 = \inf\{\eta > 0: \varphi_{1,\eta}(\xi) \leqq \varphi_2(\xi) \text{ for } \xi \leqq \xi_0\}$. If $\mu_0 > 0$, there exists $\bar{\xi} \geqq \mu_0$ such that

$$\varphi_{1,\mu_0}(\bar{\xi}) = \varphi_2(\bar{\xi}).$$

In this case we define

$$\mu_1 = \inf\{\xi \leqq \bar{\xi}: \varphi_{1,\mu_0}(\xi) = \varphi_2(\xi)\}.$$

Clearly, $\mu_1 \geqq \mu_0 > 0$. We would then have

$$\varphi_{1,\mu_0}(\xi) < \varphi_2(\xi) \quad \text{if } 0 < \xi < \mu_1, \tag{2.6}$$

$$\varphi_{1,\mu_0}(\mu_1) = \varphi_2(\mu_1) \tag{2.7}$$

but in view of (2.3c) and (2.6), $\varphi_{1,\mu_0}(\mu_1) < \varphi_2(\mu_1)$, so that (2.7) cannot hold. Thus $\mu_0 = 0$ and $\varphi_1(\xi) \leqq \varphi_2(\xi)$ for all $\xi \leqq \xi_0$. Interchanging the roles of φ_1 and φ_2 we deduce $\varphi_1(\xi) = \varphi_2(\xi)$ and the proof of (i) is complete.

(ii) Now let (φ_1, ψ_1) and (φ_2, ψ_2) be two nontrivial solutions of (2.1), (2.2). Without loss of generality, we may assume $\varphi_1(\xi) > \varphi_2(\xi)$ for $\xi > 0$. Since $\varphi_2 \not\equiv 0$, there exists $\eta > 0$ such that $\varphi_{1,\eta}(\xi) < \varphi_2(\xi)$ for $\xi > 0$. Now set

$$\eta_0 = \inf\{\eta: \varphi_{1,\eta}(\xi) < \varphi_2(\xi) \text{ for } \xi > 0\}.$$

If $\eta_0 = 0$, $\varphi_1(\xi) = \varphi_2(\xi)$ so that we may assume $\eta_0 > 0$. Then $\varphi_{1,\delta}(\xi) < \varphi_2(\xi) < \varphi_{1,\varepsilon}(\xi)$ for any δ, ε such that $0 < \varepsilon < \eta_0 < \delta$ and $\xi > 0$, and this implies that $\varphi_2(\xi) = \varphi_{1,\eta_0}(\xi)$. The corresponding result for ψ_1, ψ_2 now follows from (2.3b). □

3. Proofs of Theorems 1 and 2. The case $c < 0$. From now on we assume $pq < 1$ and look for a nontrivial solution to the integral equation for φ obtained in § 2, namely,

$$\varphi(\xi) = \int_0^\xi \left(\frac{e^{c(\xi - t)} - 1}{c}\right)\left[\int_0^t \left(\frac{e^{c(t-s)} - 1}{c}\right)\varphi(s)^q \, ds\right]^p dt. \tag{3.1}$$

To this end a fixed-point argument will be employed. We begin with the case $c < 0$, an assumption to be retained throughout this section.

Let us consider a function $h(\xi) \in \mathscr{C}^2[0, \infty) \cap \mathscr{C}^\infty(0, \infty)$ such that $h(\xi) > 0$ for $\xi > 0$ and

$$h(\xi) = \begin{cases} \xi^\alpha & \text{with } \alpha = \dfrac{2(1+p)}{1-pq} \quad \text{if } \xi \leqq 1, \\ \xi^\beta & \text{with } \beta = \dfrac{1+p}{1-pq} \quad \text{if } \xi \geqq 2. \end{cases} \tag{3.2}$$

Then (3.1) will be solved if we can find a function $\varphi(\xi)=h(\xi)f(\xi)$ such that $f(\xi)>0$ for $\xi>0$, $f\in C^2[0,\infty)$, and f is a fixed point for the operator

$$(3.3)\qquad (Tf)(\xi)=\frac{1}{h(\xi)}\int_0^{\xi}\left(\frac{e^{c(\xi-t)}-1}{c}\right)\left[\int_0^{t}\left(\frac{e^{c(t-s)}-1}{c}\right)(h(s)f(s))^q\,ds\right]^p dt$$

in the Banach space

$$(3.4)\qquad E=((C[0,\infty];\mathbb{R}),\|\ \|_\infty)$$

defined as the space of continuous functions in the Alexandroff compactified set $[0,\infty]$ corresponding to the semiaxis $[0,\infty)$, endowed with the supremum norm. We now have Lemma 3.1.

LEMMA 3.1. *T maps E into E. In particular, for any function $f\in E$,*

$$(3.5)\qquad (Tf)(0)=\frac{f(0)^{pq}}{M}\quad \textit{with } M=((1+\alpha q)(2+\alpha q))^p\alpha(\alpha-1),$$

$$(3.6)\qquad (Tf)(\infty)=\frac{f(\infty)^{pq}}{N}\quad \textit{with } N=(-c)^{1+p}(1+\beta q)^p\beta$$

where $f(\infty)=\lim_{x\to\infty}f(x)$, and α,β are as in (3.2).

Proof. It suffices to show (3.5), (3.6). When $\xi>0$ is close to zero, we have

$$\frac{e^{c(\xi-t)}-1}{c}\approx(\xi-t)\quad \text{for } 0\leqq t\leqq\xi,$$

$$f(\xi)\approx f(0)$$

so that for $\xi\approx 0$ (cf. (3.2), (3.3))

$$\begin{aligned}(Tf)(\xi)&\approx\xi^{-\alpha}\int_0^{\xi}(\xi-t)\left[\int_0^{t}(t-s)s^{\alpha q}f(0)^q\,ds\right]^p dt\\&=\xi^{-\alpha}f(0)^{pq}((1+\alpha q)(2+\alpha q))^{-p}\int_0^{\xi}(\xi-t)t^{p(\alpha q+2)}\,dt\\&=f(0)^{pq}((1+\alpha q)(2\alpha q))^{-p}(\alpha(\alpha-1))^{-1}.\end{aligned}$$

This gives (3.5). As for (3.6), we remark that since $c<0$, we have, for fixed $\xi_0>1$ and $\xi\gg\xi_0$,

$$\begin{aligned}(Tf)(\xi)&\approx\xi^{-\beta}\int_{\xi_0}^{\xi}\left(\frac{e^{c(\xi-t)}-1}{c}\right)\left[\int_{\xi_0}^{t}\left(\frac{e^{c(t-s)}-1}{c}\right)s^{\beta q}f(s)^q\,ds\right]^p dt\\&\approx\xi^{-\beta}f(\infty)^{pq}\int_{\xi_0}^{\xi}\left(\frac{-1}{c}\right)\left[\int_{\xi_0}^{t}\left(\frac{-1}{c}\right)s^{\beta q}\,ds\right]^p dt,\end{aligned}$$

whence comes the result. □

LEMMA 3.2. *T is continuous, and transforms bounded sets of E into bounded sets of E. In particular, there holds*

$$(3.7)\qquad \|Tf\|_\infty\leqq\|f\|_\infty^{pq}\|TI\|_\infty\quad \textit{where } I(\xi)=1 \textit{ for any } \xi\geqq 0.$$

The proof of this result is straightforward.

We want to show next that T transforms bounded sets into bounded and equicontinuous sets, when restricted to a suitable domain $K\subset E$, which consists of those functions $f\in E$ such that

$$(3.8a)\qquad f(\xi)\geqq 0\quad \text{for any } \xi\in[0,\infty);$$

(3.8b) $\|f\|_\infty \leqq k_1 f(0)$, $\|f\|_\infty \leqq k_2 f(\infty)$, where

$$k_1 = \max\left((M\|TI\|_\infty)^{(1/1-pq)}, 1\right), \qquad k_2 = \max\left((N\|TI\|_\infty)^{(1/1-pq)}, 1\right),$$

M and N being the constants in (3.5), (3.6), and I being the constant function in (3.7);

(3.8c) For any $\varepsilon > 0$ we have

$$(1-\varepsilon)f(0) \leqq f(x) \leqq (1+\varepsilon)f(0) \quad \text{if } x < \delta(\varepsilon),$$

$$(1-\varepsilon)f(\infty) \leqq f(x) \leqq (1+\varepsilon)f(\infty) \quad \text{if } x > N(\varepsilon)$$

where

$$\delta(\varepsilon) = \sup\left(x: (1-\varepsilon)^{1-pq} \leqq \frac{(TI)(x)}{(TI)(0)} \leqq (1+\varepsilon)^{1-pq}\right),$$

$$N(\varepsilon) = \inf\left(x: (1-\varepsilon)^{1-pq} \leqq \frac{(TI)(x)}{(TI)(\infty)} \leqq (1+\varepsilon)^{1-pq}\right).$$

We have thus defined a cone (i.e., a closed, convex set K in the Banach space E such that (i) $\lambda f \in K$ for any $\lambda \geqq 0$ if $f \in K$, and (ii) for any $f \in E$, $f \not\equiv 0$, at least one of the functions f, $(-f)$ does not lie in K). This set is nonempty, since both the trivial function and $I(x)$ in (3.7) are in K. We now have Lemma 3.3.

LEMMA 3.3. $T(K) \subset K$.

Proof. Let $f \in K$. It is clear that $(Tf)(\xi) \geqq 0$ for any $\xi \in [0, \infty)$. Furthermore, by (3.5) and (3.8b)

$$\|Tf\|_\infty \leqq \|f\|_\infty^{pq} \|TI\|_\infty \leqq k_1^{pq} f(0)^{pq} \|TI\|_\infty = k_1^{pq} M \|TI\|_\infty (Tf)(0)$$

so that we have $\|Tf\|_\infty \leqq k_1 (Tf)(0)$ as soon as we impose

$$k_1 \geqq (M\|TI\|_\infty)^{(1/1-pq)}.$$

Actually equality would suffice for our purposes at this point; however, the fact that $k_1 \geqq 1$ will play a role in a later result. Similarly, we see that $\|Tf\|_\infty \leqq k_2 (Tf)(\infty)$.

Now take $x < \delta(\varepsilon)$ with $\delta(\varepsilon)$ given in (3.8c) and use this last property applied to f to get

$$(Tf)(\xi) \leqq f(0)^{pq}(1+\varepsilon)^{pq}(TI)(\xi)$$

so that by (3.5)

$$\begin{aligned}(Tf)(\xi) &\leqq M(1+\varepsilon)^{pq}(TI)(\xi)(Tf)(0) \\ &= (1+\varepsilon)^{pq}(TI)(\xi)((TI)(0))^{-1}(Tf)(0)\end{aligned}$$

and since $(TI)(\xi)((TI)(0))^{-1} \leqq (1+\varepsilon)^{1-pq}$ in the interval under consideration, the conclusion follows. The corresponding estimate at infinity is arrived at analogously. □

LEMMA 3.4. *T transforms bounded sets in K into bounded and equicontinuous sets in K.*

Proof. The first statement in the lemma has already been listed in Lemma 3.2. On the other hand, proving equicontinuity away from $\xi = 0$ and $\xi = +\infty$ is immediate: actually, it is not necessary to restrict T to the cone K at this stage. As for the cases $\xi = 0, +\infty$, they follow at once from (3.8c) since $TK \subset K$. □

Let us summarize. So far, we have shown that T given in (3.3) is a bounded, completely continuous operator that maps the cone K defined in (3.8) into itself. By the classical Schauder Theorem (cf, for instance, [K]), T has at least a fixed point $f(\xi)$. This is not enough, however, since $f(\xi)$ may vanish identically. Our next step is then Lemma 3.5.

LEMMA 3.5. *T has a nontrivial fixed point in K.*

Proof. For $n = 1, 2, \cdots$ we define

$$T_n(f) = T\left(f + \frac{I}{n}\right)$$

where $I(x)$ is given in (3.7). Now, since $k_1 \geqq 1$ and $k_2 \geqq 1$, the function $f + I/n$ belongs to K provided that $f \in K$, and it is easy to see that for any fixed n, T_n is a nonnegative, bounded, and completely continuous operator from K into K. Furthermore, since

$$\| TI \|_\infty \geqq \min\left(TI(0), TI(\infty)\right) = \min\left(\frac{1}{M}, \frac{1}{N}\right) \equiv a_0 > 0,$$

we have

$$\| T_n f \|_\infty \geqq \left\| T\frac{I}{n} \right\|_\infty \geqq \frac{a_0}{n^{pq}} > 0. \tag{3.9}$$

Then a variant of Schauder's theorem (cf. [K, p. 242]) states that for any $r_0 > 0$, there exist $\lambda_n > 0$ and $g_n \in K$ such that

$$T_n g_n = \lambda_n g_n, \tag{3.10}$$

$$\| g_n \|_\infty = r_0. \tag{3.11}$$

Write $\mu_n = (\lambda_n)^{-1}$. It then follows from (3.9) that

$$g_n(\xi) = \mu_n T_n g_n(\xi) = \mu_n T\left(g_n(\xi) + \frac{I(\xi)}{n}\right) \geqq \mu_n T\left(\frac{I(\xi)}{n}\right)$$

whence

$$\begin{aligned} g_n(\xi) = \mu_n T_n g_n(\xi) &= \mu_n T\left(g_n(\xi) + \frac{I(\xi)}{n}\right) \geqq \mu_n T\left(g_n(\xi)\right) \\ &\geqq \mu_n T\left(\mu_n T\left(\frac{I(\xi)}{n}\right)\right) = \mu_n^{1+pq} n^{-(pq)^2} T^2 I(\xi). \end{aligned} \tag{3.12}$$

Iterating (3.12), we deduce that for any $m = 1, 2, \cdots$

$$g_n(\xi) \geqq (\mu_n^{1+pq+\cdots+(pq)^m}) n^{-pq^{(m+1)}} T^{m+1} I(\xi)$$

so that

$$r_0 = \| g_n \|_\infty \geqq (\mu_n^{1+pq+\cdots+(pq)^m}) n^{-pq^{(m+1)}} \| T^{m+1} I \|_\infty. \tag{3.13}$$

On the other hand, by (3.5)

$$\| T^2 I \|_\infty = \| T(TI) \|_\infty \geqq T(TI(0)) = T\left(\frac{1}{M}\right) = (M^{1+pq})^{-1}.$$

Thus for any $m \geqq 1$

$$\| T^{m+1} I \|_\infty \geqq (M^{1+pq+\cdots+(pq)^m})^{-1}. \tag{3.14}$$

Putting together (3.13) and (3.14) we get

$$r_0 = \|g_n\|_\infty \geqq (M^{1+pq+\cdots+(pq)^m})^{-1} n^{-pq^{(m+1)}} (\mu_n^{1+pq+\cdots+(pq)^m}). \tag{3.15}$$

We now let $m \to \infty$ in (3.15) for fixed n. As $pq < 1$, this gives

$$r_0 \geqq (M^{(1/(1-pq))})^{-1} \mu_n^{(1/(1-pq))}. \tag{3.16}$$

The next step consists of passing to the limit in (3.10). By (3.11), the sequence $\{g_n + I/n\}$, $n = 1, 2, \cdots$ is bounded. As T is completely continuous, we have that, up to a subsequence also labelled with n,

$$T\left(g_n + \frac{I}{n}\right) \to g \quad \text{as } n \to \infty \tag{3.17}$$

for some $g \in K$ with $0 < \|g\|_\infty < \infty$. On the other hand, by (3.16) and again up to a subsequence, we get

$$\mu_n \to \mu_0 < \infty \quad \text{as } n \to \infty. \tag{3.18}$$

Now from (3.10), (3.17), and (3.18) we conclude that, along a suitable subsequence,

$$g_n \to \mu_0 g \quad \text{as } n \to \infty. \tag{3.19}$$

Therefore, using (3.17) and (3.19), we get

$$g = \lim_{n\to\infty} T\left(g_n + \frac{I}{n}\right) = \lim_{n\to\infty} T(g_n) = T(\mu_0 g),$$

i.e.,

$$g = T(\mu_0 g) = \mu_0^{pq} T(g)$$

and then the function $f = \mu_0^{-(pq/(1-pq))} g$ is the required fixed point, since it satisfies $f = Tf$. □

Note that, since $\|f\|_\infty \neq 0$, it follows from (3.8c) that $f(0)$ and $f(\infty)$ are both positive. In particular, from (3.5), (3.6) we conclude that $f(0)^{1-pq} = M^{-1}$, $f(\infty)^{1-pq} = N^{-1}$ with M, N as in Lemma 3.1. This gives the asymptotic estimates corresponding to φ (1.4), (1.5) of Theorem 2. Those corresponding to ψ follow, for instance, from (2.3b). □

4. The case $c > 0$. Our next step consists in obtaining the corresponding results for expanding waves, i.e., for the case $c > 0$. Since the arguments very much parallel those in § 3, we just stress the new relevant points and sketch the rest of the proof. To begin with, instead of $h(\xi)$ in (3.2) we consider three positive functions $z_i(\xi) \in \mathscr{C}^2[0, \infty)$, $i = 1, 2, 3$ such that $z_i(\xi) = \xi^\alpha$ with $\alpha = 2(1+p)/(1-pq)$ if $\xi \leqq 1$ and

$$z_i(\xi) = \begin{cases} e^{c\xi} & \text{if } \xi \geqq 2, \quad i = 1, \\ \xi e^{c\xi} & \text{if } \xi \geqq 2, \quad i = 2, \\ e^{c\xi} & \text{if } \xi \geqq 2, \quad i = 3. \end{cases} \tag{4.1}$$

We now define nonnegative operators $T_{i,q}$ acting on E (cf. (3.4)) as follows:

$$(T_{i,q} f)(\xi) = \frac{1}{z_i(\xi)} \int_0^\xi \left(\frac{e^{c(\xi-t)} - 1}{c}\right) \left[\int_0^t \left(\frac{e^{c(t-s)} - 1}{c}\right) z_i^q(s) f(s)^q \, ds\right]^p dt$$

where $i = 1$ if $p < 1$, $i = 2$ if $p = 1$, and $i = 3$ if $p > 1$. Then we have Lemma 4.1.

LEMMA 4.1. *T_i maps E into E for $i=1,2,3$. In particular, for any $f\in E$,*

$$(T_{i,q}f)(0)=\frac{f(0)^{pq}}{M} \quad \textit{with } M \textit{ given in } (3.5) \quad \textit{and} \quad i=1,2,3, \tag{4.2}$$

$$(T_{1,q}f)(\infty)=f(\infty)\frac{1}{(c^2(1-q))^p c^2(1-p)} \quad \textit{if } q<1, \tag{4.3a}$$

$$(T_{1,q}f)(\infty)=f(\infty)^p\frac{1}{c^{p+1}}\int_0^{+\infty} s^p\, e^{c(p-1)s}\,ds \quad \textit{if } q=1, \tag{4.3b}$$

$$(T_{1,q}f)(\infty)=f(\infty)^{pq}\frac{1}{(c^2(q-1)q)^p c^2(1-pq)} \quad \textit{if } q>1. \tag{4.3c}$$

Proof. Condition (4.2) has already been obtained in (3.5). As for (4.3), we note that since

$$\int_0^{\xi}\left(\frac{e^{c(\xi-t)}-1}{c}\right)e^{cqt}\,dt=\frac{e^{cq\xi}}{c^2q(q-1)}-\frac{e^{c\xi}}{c^2(q-1)}+\frac{1}{c^2q},$$

we have for $\xi\gg 0$

$$\left[\int_0^{\xi}\left(\frac{e^{c(\xi-t)}-1}{c}\right)e^{cqt}\,dt\right]^p\approx\begin{cases}\dfrac{e^{cp\xi}}{(c^2(1-q))^p} & \text{if } q<1,\\[2ex] \dfrac{\xi^p\, e^{cp\xi}}{c^p} & \text{if } q=1,\\[2ex] \dfrac{e^{cpq\xi}}{(c^2q(q-1))^p} & \text{if } q>1,\end{cases} \tag{4.4}$$

and then (4.3) follows by elementary calculus. □

To conclude the proof of Theorem 2, it suffices to repeat the fixed-point arguments in § 3 with some obvious modifications. In particular, for $p<1$ operators $T_{1,q}$ are to be considered as acting on the cones $K_{i,q}$, obtained by replacing (TI) by $(T_{1,q}I)$ in (3.8). □

Proof of Theorem 3. It will suffice to obtain (1.14) and (1.16), since the corresponding results for $v_c(x,t)$ are similar. To proceed, we recall that in Theorem 2 we have shown that

$$\varphi(\xi)=\xi^{\alpha}f(\xi) \quad \text{for } \xi>0, \quad \xi\approx 0 \tag{4.5}$$

where

$$f(\xi)=(Tf)(\xi), \tag{4.6}$$

T being the operator defined in (3.3), (3.4). Since $f(\xi)\in\mathscr{C}^2$, we have that

$$f(\xi)=f(0)+f'(0)\xi+O(\xi^2) \quad \text{as } \xi\approx 0, \tag{4.7}$$

whence

$$A_1=f'(0)=\lim_{\xi\to 0^+}\left(\frac{1}{\xi}((Tf)(\xi)-A)\right). \tag{4.8}$$

To compute A_1 we expand the quantity in the braces above up to second-order terms and then pass to the limit. For $t \approx 0$, we compute

$$\int_0^t \left(\frac{e^{c(t-s)}-1}{c}\right) s^{\alpha q} f(s)^q \, ds$$

$$\approx \int_0^t \left((t-s)+\frac{c}{2}(t-s)^2\right) s^{\alpha q} A^q \left(1+\frac{A_1 s}{A}\right)^q ds$$

$$\approx \int_0^t \left((t-s)+\frac{c}{2}(t-s)^2\right) s^{\alpha q} (A^q + qA^{q-1}A_1)\, ds$$

$$\approx \frac{A^q t^\beta}{\beta(\beta-1)} + \frac{t^{\beta+1}A^{q-1}}{\beta(\beta+1)}\left(\frac{cA}{\beta-1}+qA_1\right)$$

where β is given in (1.4). It then follows that

$$\int_0^\xi \left(\frac{e^{c(\xi-t)}-1}{c}\right)\left[\int_0^t \left(\frac{e^{c(t-s)}-1}{c}\right) s^{\alpha q} f(s)^q \, ds\right]^p dt$$

$$\approx \frac{A^{pq}\xi^\alpha}{(\beta(\beta-1))^p \alpha(\alpha-1)} + \frac{A^{pq}\xi^{\alpha+1}}{(\beta(\beta-1))^p \alpha(\alpha+1)}$$

$$\cdot \left(\frac{c(p(\alpha-1)+\beta+1)}{(\alpha-1)(\beta+1)} + \frac{pqA^{-1}A_1(\beta-1)}{\beta+1}\right)$$

so that, taking into account (1.4), we conclude that for $\xi > 0$, $\xi \approx 0$,

$$\frac{1}{\xi}((Tf)(\xi)-A) \approx \frac{Ac(p(\alpha-1)+\beta+1)+pqA_1(\alpha-1)(\beta-1)}{(\alpha+1)(\beta+1)}. \tag{4.9}$$

The results follow from (4.8) and (4.9). □

We conclude with a few remarks. First we observe that, for the waves (u_c, v_c) obtained previously, we have

$$\lim_{ct-x\to 0^+} \left(\frac{\partial u_c(ct-x)/\partial t}{\partial u_c(ct-x)/\partial x}\right) = -c, \tag{4.10}$$

which is the equation satisfied at the front of u_c. Indeed, a similar result holds for v_c. Note that relation (4.10) can be arrived at in a formal way as follows. For $\varepsilon > 0$, write $x_\varepsilon(t)$ to denote the level line $u(x_\varepsilon(t), t) = \varepsilon$. Then differentiating with respect to t yields

$$\frac{\partial u}{\partial x}\frac{dx_\varepsilon}{dt} + \frac{\partial u}{\partial t} = 0,$$

which, if we let $\varepsilon \to 0$, suggest the following equation for the level line $x_0(t)$, where $u(x_0(t), t) = 0$:

$$\frac{dx_0}{dt} = -\frac{\partial u/\partial t}{\partial u/\partial x}$$

where the quantity on the right is computed at the curve $x_0(t)$. This is precisely (4.10) in the case of waves (u_c, v_c).

Next we consider the limit case $pq = 1$. We already know that there are no FTW's in this case. However, travelling waves exist, although they are not unique. More precisely, we have Lemma 4.2.

LEMMA 4.2. *Assume* $pq=1$. *Then for any real* c *there exist* $\theta>0$, $\chi>\max(0, c)$ *such that, for any* $P>0$,

$$\varphi(\xi)=Pe^{\theta\xi}, \qquad \psi(\xi)=\frac{p^q e^{\chi\xi}}{\chi(\chi-c)}$$

is a monoparametric family of solutions to (1.7) *satisfying*

$$\varphi(\xi), \psi(\xi)\to\infty \quad \text{as } \xi\to\infty,$$

$$\varphi(\xi), \psi(\xi)\to 0 \quad \text{as } \xi\to-\infty.$$

Proof. We just try $\varphi=Pe^{\theta\xi}$, $\psi=Qe^{\chi\xi}$ in (1.7). This leads to

$$\begin{aligned} P\theta(\theta-c)\,e^{\theta\xi}&=Q^p\,e^{\chi p\xi},\\ Q\chi(\chi-c)\,e^{\chi\xi}&=P^q\,e^{\theta q\xi} \end{aligned} \tag{4.11}$$

where we impose

$$\theta=\chi p, \qquad \chi=\theta q. \tag{4.12}$$

Note that this yields no further information on θ or χ, since $pq=1$. Matching the corresponding coefficients in (4.11), we get $P\theta(\theta-c)=Q^p$, $Q\chi(\chi-c)=P^q$ whence

$$P\theta(\theta-c)=\left(\frac{P^q}{\chi(\chi-c)}\right)^p=\frac{P}{(\chi(\chi-c))^p}$$

and using (4.12) we deduce

$$\chi p(\chi p-c)(\chi(\chi-c))^p=1.$$

When $p=q=1$ this reads $\chi^2(\chi-c)^2=1$, which always has a positive solution for $\chi\geqq\max(0, c)$. As to the general case, we may assume, for instance, that $p>1$, and then $F_p(\chi)=\chi p(\chi p-c)(\chi(\chi-c))^p$ also has a positive root for $\chi\geqq\max(0, c)$. The corresponding values of θ are then obtained by (4.12), and for fixed $P>0$ the associated $Q=Q(P)$ is determined from (4.11). □

As a final remark, we observe that the methods in this paper apply to other types of semilinear diffusion-absorption systems. For instance, consider the following example:

$$\begin{aligned} &u_t-u_{xx}+(uv)^p=0,\\ &v_t-v_{xx}+(uv)^q=0,\\ &-\infty<x<\infty,\ t>0, \qquad p, q>0. \end{aligned} \tag{4.13}$$

Searching for waves $u_c(x, t)=\varphi(ct-x)$, $v_c(x, t)=\varphi(ct-x)$ in (4.13) we are led to the ODE system

$$\varphi''=c\varphi'+(\varphi\psi)^p, \qquad \psi''=c\psi'+(\varphi\psi)^q. \tag{4.14}$$

Now the analogue to (1.3), i.e., the condition for FTW's to exist, is

$$p+q<1.$$

As before, formal perturbation methods suggest the asymptotic behaviour of the waves.

For instance, in the case $c<0$, we have

$$\varphi(\xi)\approx\begin{cases}A_1\xi^{\alpha_1} & \text{with } \alpha_1=\dfrac{2(1-q+p)}{1-(p+q)} \quad \text{if } \xi\approx 0,\\ A_2\xi^{\alpha_1/2} & \text{if } \xi\gg 0,\end{cases}$$

$$\psi(\xi)\approx\begin{cases}B_1\xi^{\beta_1} & \text{with } \beta_1=\dfrac{2(1-p+q)}{1-(p+q)} \quad \text{if } \xi\approx 0,\\ B_2\xi^{\beta_1/2} & \text{if } \xi\gg 0\end{cases}$$

where A_1, A_2, B_1, B_2 are positive constants depending on p and q.

Results corresponding to Theorems 1-3 can now be proved with some minor modifications. We just mention that, because of the structure of (4.14), it is convenient to get the fixed-point argument for the variable $z(\xi)=\varphi(\xi)\psi(\xi)$, for which the formula analogous to (2.4) reads

$$z(\xi)=\left[\int_0^\xi\left(\frac{e^{c(\xi-t)}-1}{c}\right)z(t)^p\,dt\right]\left[\int_0^\xi\left(\frac{e^{c(\xi-t)}-1}{c}\right)z(t)^q\,dt\right].$$

REFERENCES

[A] R. ARIS, *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts*, Vols. I and II, Clarendon Press, Oxford, 1975.

[AW] D. G. ARONSON AND H. F. WEINBERGER, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. in Math., 30 (1978), pp. 33-76.

[BNS] H. BERESTYCKI, B. NICOLAENKO, AND B. SCHEURER, *Travelling wave solutions to reaction diffusion systems modelling combustion*, in Proc. Conference on Partial Differential Equations, Durham, NC, 1982, American Mathematical Society, Providence, RI.

[CL] P. CLAVIN AND A. LIÑÁN, *Theory of gaseous combustion*, in Nonequilibrium Cooperative Phenomena in Physics and Related Fields, M. G. Velarde, ed., Plenum Press, New York, 1984.

[F] P. C. FIFE, *Mathematical Aspects of Reacting and Diffusing Systems*, Lecture Notes in Biomath. 28, Springer-Verlag, Berlin, New York, 1970.

[FM] P. C. FIFE AND J. B. MCLEOD, *A phase plane discussion of convergence to travelling fronts for nonlinear diffusion*, Arch. Rational Mech. Anal., 75 (1981), pp. 281-314.

[H] S. HEINZE, *Travelling waves in combustion processes with complex chemical networks*, Trans. Amer. Math. Soc., 304 (1987), pp. 405-416.

[HV] M. A. HERRERO AND J. L. VÁZQUEZ, *Thermal waves in absorbing media*, J. Differential Equations, 74 (1988), pp. 218-233.

[K] M. A. KRASNOSEL'SKII, *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon Press, Oxford, 1964.

[M] L. K. MARTINSON, *The finite velocity of propagation of thermal perturbations in media with constant thermal conductivity*, Zh. Vychisl. Mat. i Mat. Fiz., 16 (1976), pp. 1233-1241.

[PP] L. D. POKROVSKII AND S. N. PARANENKO, *Conditions for space localization of the solutions of the non-linear equation of heat conduction*, U.S.S.R. Comput. Math. and Math. Phys., 22 (1982), pp. 264-269.

[RK] PH. ROSENAU AND S. KAMIN, *Thermal waves in an absorbing and convecting medium*, Phys. D, 8, (1983), pp. 273-283.

[S] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, New York, 1982.

[T] D. TERMAN, *Travelling wave solutions arising from a combustion model*, IMA Preprint 216, University of Minnesota, Minneapolis, MN, 1986.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 137-153, January 1990
© 1990 Society for Industrial and Applied Mathematics
008

SPATIALLY OSCILLATORY STEADY STATES OF TUBULAR CHEMICAL REACTORS*

ROGER ALEXANDER†

Abstract. The equations governing the nonadiabatic tubular chemical reactor have as many low-conversion steady-state solutions as are wanted, if the coefficient of heat transfer from reactor to cooling jacket is sufficiently large, and if the activation energy is large. These steady states exhibit no reaction zone: temperature and reactant concentration do not deviate much from their inlet values. The temperature profiles are oscillatory in these steady states; the most oscillatory profile can be computed by the method of averaging.

Key words. chemical reactors, multiple steady states, averaging method

AMS(MOS) subject classifications. 80A32, 34B15

1. Introduction. In this paper we show that the axial dispersion model of the nonadiabatic tubular reactor can have arbitrarily many steady states, provided that two constants occurring in the governing equations are sufficiently large.

The steady states of the reactor are solutions of a boundary value problem for the temperature T and reacting species concentration C. In these equations the reactor length has been normalized to unity. Constants H, M, B, D, β, γ are explained below. The equations are:

$$
\begin{aligned}
&\frac{1}{H}T''-T'-\beta(T-1)+BDC\,e^{-\gamma/T}=0, && 0<x<1,\\
&T'-H(T-1)=0, && x=0,\\
&T'=0, && x=1,\\
&\frac{1}{M}C''-C'-D\,e^{-\gamma/T}C=0, && 0<x<1,\\
&C'-M(C-1)=0, && x=0,\\
&C'=0, && x=1.
\end{aligned}
\tag{1.1}
$$

Consult [VA] for a detailed derivation, or [A2] for an explanatory sketch; here we merely indicate the meaning of the parameters:

H, M—Péclet numbers for heat and mass, respectively,

β—coefficient of heat transfer between reactor tube and cooling jacket,

B—heat release of chemical reaction,

D—Damköhler number,

γ—activation energy.

As far as the author knows, no one has proved that (1.1) can have more than three solutions. However, formal asymptotic methods and numerical bifurcation analysis have made it "well known" that there can be up to seven steady states when the activation energy is large. This result, and the difficulties the equations pose for a rigorous analysis, are discussed in the survey [A2].

* Received by the editors June 6, 1988; accepted for publication February 27, 1989. This research was supported by Air Force Office of Scientific Research grant 84-0252 and Air Force Systems Command grant 88-0031. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation therein.

† Department of Mathematics, Iowa State University, Ames, Iowa 50011.

Kapila and Poore [KP] have constructed formal asymptotic solutions to (1.1). We are interested here in *one* of the types of solution described by them, the "low-conversion" steady state. Such a solution exhibits no "reaction zone"; instead T and C remain close to their inlet values throughout the reactor.

In this paper we show that (1.1) can have an arbitrary number of low-conversion steady-state solutions, provided that the reactor parameters are in a suitable range. These solutions exhibit temperature profiles in the form of oscillations that grow slowly in amplitude as the reactor is traversed from inlet to outlet. Note that these solutions occur in addition to the six other types of steady states, identified by Kapila and Poore, that may be present.

This paper extends previous work of the author [A1] in two crucial respects. First, in [A1] multiple solutions are found for an approximating equation—what is called here the "decoupled equation" for the temperature.

Here we derive the *same* multiplicity results for the original equation (1.1) as for the decoupled equation.

The second improvement over [A1] is the use of the method of averaging to obtain a more precise count of the number of solutions. In the previous paper, a simple differential inequality is used to give a lower bound on solution multiplicity.

Formally, the limit $\gamma \to \infty$ yields the "nonreacting" solution: $T = C = 1$ everywhere. We take γ finite but large, and use it as a microscope to find "low-conversion" solutions—solutions with T and C near 1 throughout the reactor.

Into (1.1) we substitute

$$T = 1 + \gamma^{-1} y, \quad C = 1 + \gamma^{-1} z, \quad D = \gamma^{-1} \lambda e^{\gamma}$$

to obtain an equivalent boundary problem for (y, z):

$$\frac{1}{H} y'' - y' - \beta y + B\lambda (1 + \gamma^{-1} z) \exp\left[\frac{y'}{1 + \gamma^{-1} y}\right] = 0, \qquad 0 < x < 1, \tag{1.2}$$

$$y' - Hy = 0 \quad \text{at } x = 0, \tag{1.2$_0$}$$

$$y' = 0 \quad \text{at } x = 1, \tag{1.2$_1$}$$

$$\frac{1}{M} z'' - z' - \gamma^{-1} \lambda \exp\left[\frac{y}{1 + \gamma^{-1} y}\right] z = \lambda \exp\left[\frac{y}{1 + \gamma^{-1} y}\right], \qquad 0 < x < 1, \tag{1.3}$$

$$z' - Mz = 0 \quad \text{at } x = 0, \tag{1.3$_0$}$$

$$z' = 0 \quad \text{at } x = 1. \tag{1.3$_1$}$$

In (1.3), z is the solution of a linear boundary problem with coefficients depending on y. On the other hand, we expect z to influence y only weakly, for it appears multiplied by γ^{-1} in the equation for y, and γ is large.

It is worth emphasizing that (1.2)-(1.3) are *equivalent* to the original system of (1.1)—they do not represent the first term in an expansion in inverse powers of γ.

Replace the coefficient $(1 + \gamma^{-1} z)$ in (1.2) by 1 to obtain the "decoupled equation" for y:

$$\frac{1}{H} y'' - y' - \beta y + B\lambda \exp\left[\frac{y}{1 + \gamma^{-1} y}\right] = 0, \qquad 0 < x < 1,$$

$$y' - Hy = 0 \quad \text{at } x = 0, \tag{1.4}$$

$$y' = 0 \quad \text{at } x = 1.$$

In § 2 we study this problem. We show that it has many solutions when β is large and $\gamma^{-1}\beta$ is small. We establish the oscillatory character of those solutions and derive bounds on them. Phase-plane methods have been used before to investigate multiple steady states (see [MA]). In § 3 we return to the coupled problem, and show that to each solution $\tilde{y}$ of (1.4) corresponds a solution (y, z) of (1.2)-(1.3) with y near $\tilde{y}$; moreover, distinct solutions of (1.4) yield distinct solutions of (1.2)-(1.3). Section 4 summarizes some further points and unanswered questions. Some calculations are relegated to the Appendices.

2. The decoupled y-equation. Begin by writing (1.4) as a boundary value problem for a first-order system. Let

$$y_1 = y, \qquad y_2 = y'.$$

This gives

$$y_1' = y_2, \qquad y_2' = H\left\{y_2 + \beta y_1 - B\lambda \exp\left[\frac{y_1}{1+\gamma^{-1}y_1}\right]\right\}, \tag{2.1}$$

and the boundary conditions

$$y_2(0) - Hy_1(0) = 0, \qquad y_2(1) = 0. \tag{2.2}$$

Let the parameters H, B, λ be fixed in what follows. It is shown in Appendix A that if $\beta/B\lambda > e$ and $\gamma^{-1}\beta$ is small, then the system of (2.1) has three critical points $(y_1, y_2) = (\alpha_j, 0)$, $j = 0, 1, 2$:

$$\alpha_0 = \frac{B\lambda}{\beta} + O(\beta^{-2}),$$

$$\alpha = \alpha_1 = \log\frac{\beta}{B\lambda} + \log\log\frac{\beta}{B\lambda} + o(1),$$

$$\alpha_2 = \frac{B\lambda}{\beta}\, e^{\gamma}(1 + O(\gamma^2 e^{-\gamma})).$$

The critical points $(\alpha_0, 0)$ and $(\alpha_2, 0)$ are always saddles. We will write α for α_1—this is the pivot of our analysis. Appendix A shows as well that the critical point $(\alpha, 0)$ is an unstable spiral point provided

$$\frac{\alpha}{(1+\gamma^{-1}\alpha)^2} > 1 + \frac{H}{4\beta} \qquad (\alpha = \alpha_1).$$

This condition holds when β^{-1} and $\gamma^{-1}\beta$ are small enough, for any fixed choice of H, B, and λ.

Let us write $y(x; \eta) := (y_1(x; \eta), y_2(x; \eta))^T$ for the solution of the differential equations (2.1) subject to the initial conditions

$$y_1(0) = \eta, \qquad y_2(0) = H\eta.$$

This y satisfies the first boundary condition of (2.2). It remains to determine η so that the second boundary condition of (2.2), $y_2(1; \eta) = 0$, is satisfied. We show that there are many such η.

Keeping in mind the geometry of trajectories near the spiral point, use the rule

$$\hat{\theta}(x; \eta) = \arctan\frac{y_2(x; \eta)}{y_1(x; \eta) - \alpha}$$

to determine $\hat{\theta}$ depending continuously on both variables $x \geqq 0$, $\eta \geqq 0$, understanding that

$$\pi \geqq \hat{\theta}(0; \eta) \geqq 0.$$

Solutions of the decoupled equation and boundary conditions correspond to values of η for which

$$\hat{\theta}(1; \eta) = k\pi, \qquad k \in \mathbf{Z}.$$

It is easy to see from the vector field that

$$\hat{\theta}(1; 0) > \pi. \tag{2.3}$$

We will establish Lemma 2.1.

LEMMA 2.1. *There are numbers $\bar{\eta}$ and Y with $\alpha < \bar{\eta} < Y = \alpha + \log\log\sqrt{\beta/B\lambda} + o(1)$ such that*

$$\hat{\theta}(1; \bar{\eta}) > -\pi.$$

The bulk of this section is devoted to the proof of the next lemma.

LEMMA 2.2. *There is a number η^* with $0 < \eta^* < \bar{\eta}$ such that*

$$\hat{\theta}(1; \eta^*) < -\omega + O(1), \qquad \omega \sim (\beta \log \beta)^{1/2}.$$

The multiplicity result then follows directly from (2.3) and Lemmata 2.1 and 2.2 by an application of the continuous dependence of solutions on initial conditions and the Intermediate Value Theorem.

THEOREM. *Let N be the greatest integer less than ω/π. Then for each integer $k = 1, 0, -1, \cdots, -N$ there is a number η'_k satisfying*

$$0 < \eta'_k < \eta^*, \qquad \hat{\theta}(1; \eta'_k) = k\pi;$$

for each integer $k = -1, -2, \cdots, -N$ there is a number η''_k satisfying

$$\eta^* < \eta''_k < \bar{\eta}, \qquad \hat{\theta}(1; \eta''_k) = k\pi.$$

Since ω may be made arbitrarily large by making β^{-1} and $\gamma^{-1}\beta$ small, the decoupled equation (2.1) with boundary conditions (2.2) can have an arbitrary, finite number of solutions.

This is the argument of [A1], refined so as not to give away half the solutions. The proof of Lemma 2.1 yields the following corollary.

COROLLARY TO LEMMA 2.1. *There is a constant independent of β, γ such that*

$$\max_{\substack{0 \leqq x \leqq 1 \\ 0 \leqq \eta \leqq \bar{\eta}}} |y_1(x; \eta)| \leqq \text{const } Y.$$

This bound will be useful in the construction of solutions of the coupled equations in § 3.

Finally, we note that [A1] works with the simplified nonlinear reaction rate e^y instead of the "exact" form $\exp[y/(1+\gamma^{-1}y)]$. This makes no essential difference because γ is large and the argument y is restricted to the range $0 \leqq y \leqq Y$.

Proof of Lemma 2.1 *and its corollary.* We will determine $Y > \alpha$ such that the right branch of the stable manifold of the saddle $(\alpha_0, 0)$ crosses the y_1-axis between $(\alpha, 0)$ and $(Y, 0)$. Inspection of the vector field and uniqueness of solutions then shows that $y_2(x; Y)$ decreases monotonically to zero and remains negative thereafter. This proves the result.

To determine Y, consider the conservative system

$$y_1' = y_2, \qquad y_2' = H\beta\left(y_1 - \frac{B\lambda}{\beta}\exp\left[\frac{y_1}{1+\gamma^{-1}y_1}\right]\right). \tag{2.4}$$

Define the potential V by

$$-V'(y) = H\beta\left(y - \frac{B\lambda}{\beta}\exp\left[\frac{y}{1+\gamma^{-1}y}\right]\right), \qquad V(0) = 0.$$

Consider the Hamiltonian function

$$K(y_1, y_2) = \tfrac{1}{2}y_2^2 + V(y_1).$$

Differentiation along a solution of (2.1) yields

$$\frac{d}{dx}K(y_1, y_2) = Hy_2^2 \geqq 0$$

with the inequality being strict everywhere but on the y_1-axis.

The level curve

$$K(y_1, y_2) = K(\alpha_0, 0) \tag{2.5}$$

is a saddle-loop for the conservative system (2.4). We have Proposition 2.1.

PROPOSITION 2.1. *K is a Lyapunov function for the system* (2.1), *run backward in x. Every point inside the saddle-loop* (2.5) *tends under* (2.1) *to the spiral* $(\alpha, 0)$ *as* $x \to -\infty$. *In particular, the right branch of the stable manifold of the saddle* $(\alpha_0, 0)$ *spirals into* $(\alpha, 0)$ *as* $x \to -\infty$.

Proof [H]. If we call $(Y, 0)$ the vertex of the saddle-loop equation (2.5) for the conservative system, then the right branch of the stable manifold of the saddle $(\alpha_0, 0)$ first crosses the y_1 axis for negative x at a point between $(\alpha, 0)$ and $(Y, 0)$. The approximation to Y is computed in Appendix B. There we show that

$$\begin{aligned} Y &= \log\frac{\beta}{B\lambda} + 2\log\log\frac{\beta}{B\lambda} - \log 2 + o(1) \\ &= \alpha + \log\log\sqrt{\beta/B\lambda} + o(1). \end{aligned}$$

If we follow the right branch of the stable manifold of the saddle backward in x beyond its crossing of the y_1-axis we find its first crossing of the line of initial conditions at a point we will call $(\bar{\eta}, H\bar{\eta})$. Note that $\bar{\eta} < Y$.

Proof of Lemma 2.2. To begin, change to local coordinates about the spiral point $(\alpha, 0)$. Appendix C gives the boundary problem (2.1), (2.2) in terms of new variables (u_1, u_2):

$$\begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \begin{bmatrix} H/2 & \omega \\ -\omega & H/2 \end{bmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + Q\omega N(u_1)\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{2.6}$$

$$u_2(0) = \frac{H}{2\omega}u_1(0) + \frac{H\alpha}{\omega}, \qquad u_2(1) = -\frac{H}{2\omega}u_1(1). \tag{2.7}$$

In these equations the parameters ω and Q satisfy

$$\omega \sim \sqrt{\beta\log\beta}, \qquad Q \sim 1$$

when β^{-1} and $\gamma^{-1}\beta$ are small. The nonlinear function N satisfies $N(0) = N'(0) = 0$.

What makes this problem interesting is that the coefficient of the nonlinear term is large while the neighborhood of the origin in which we need a good approximate

solution is small: we will see that unlimited multiplicity of solutions is tied to $\omega \to \infty$, but ω is the coefficient of the nonlinear term of (2.6).

Introduce the polar coordinates

$$u_1 = r\cos\theta, \qquad u_2 = r\sin\theta.$$

Equation (2.6) becomes

$$\begin{aligned} r' &= \frac{H}{2} r - Q\omega \sin\theta\, N(r\cos\theta), \\ \theta' &= -\omega\left(1 + \frac{Q\cos\theta}{r} N(r\cos\theta)\right). \end{aligned} \tag{2.8}$$

The boundary conditions, expressed in terms of the angle

$$\psi = \tan^{-1}(H/2\omega),$$

are

$$\begin{aligned} r(0) &= H\frac{\alpha}{\omega}\csc(\theta(0)-\psi)\Big/\sqrt{1+\left(\frac{H}{2\omega}\right)^2}, \\ \tan\theta(1) &= -\tan\psi = -\frac{H}{2\omega}. \end{aligned} \tag{2.9}$$

We seek the initial point of the trajectory that wraps around the spiral point the maximum number of times, starting from the line of initial conditions (2.9). Change dependent and independent variables by

$$t = \omega x; \qquad \text{write "}\cdot\text{" for } \frac{d}{dt};$$

$$R(t) = \omega^{1/2} r(t/\omega), \qquad \Theta(t) = \theta(t/\omega).$$

This change of variables makes (2.8) into

$$\begin{aligned} \dot{R} &= \frac{H}{2\omega} R - Q\omega^{1/2} \sin\Theta\, N(\omega^{-1/2} R\cos\Theta), \\ \dot{\Theta} &= -1 - \frac{\omega^{1/2} Q\cos\Theta}{R} N(\omega^{-1/2} R\cos\Theta). \end{aligned}$$

Finally, the substitution $\Theta(t) = \varphi(t) - t$ puts the equation into a form suitable for averaging:

$$\begin{aligned} \dot{R} &= \frac{H}{2\omega} R - \omega^{1/2} Q\sin(\varphi - t) N(\omega^{-1/2} R\cos(\varphi - t)), \\ \dot{\varphi} &= -\omega^{1/2}\frac{Q}{R}\cos(\varphi - t) N(\omega^{-1/2} R\cos(\varphi - t)). \end{aligned} \tag{2.10}$$

The boundary conditions are

$$\begin{aligned} R(0) &= \frac{H}{\sqrt{1+\left(\frac{H}{2\omega}\right)^2}} \frac{\alpha}{\omega^{1/2}} \csc(\varphi(0) - \psi), \\ &\tan(\varphi(\omega) - \omega) + \tan\psi = 0. \end{aligned} \tag{2.11}$$

For the averaging computation, we expand (2.10) in powers of $\omega^{-1/2}$:

$$(\dot{R}, \dot{\varphi})^T = \omega^{-1/2} f(t, R, \varphi) + \omega^{-1} g(t, R, \varphi) + \omega^{-3/2} p(t, R, \varphi, \omega^{-1/2}).$$

The analytical forms of f, g, p are computed in Appendix D. Each of f, g, p is 2π-periodic in t, and the time-average of f is zero:

$$f^0(R, \varphi) = \int_0^{2\pi} f(t, R, \varphi)\, dt = 0.$$

This fortunate cancellation allows us to compute an approximate solution valid (i.e., with error $O(\omega^{-1/2})$) for $0 \leqq t \leqq \omega$, that is, for the entire interval $0 \leqq x \leqq 1$. This follows from Theorem 3.9.1 of [SV]. It is shown in Appendix D that the solution of (2.10) with the initial condition (2.11) is approximated to within an error of $O(\omega^{-1/2})$ for $0 \leqq t \leqq \omega$ by the solution of

$$\begin{aligned} &\frac{d\bar{R}}{dt} = \omega^{-1} \frac{H}{2} \bar{R}, \qquad && \bar{R}(0) = R(0), \\ &\frac{d\bar{\varphi}}{dt} = \omega^{-1} E \bar{R}^2, && \bar{\varphi}(0) = \varphi(0). \end{aligned} \tag{2.12}$$

(E is a constant defined by (D4).)

Solve these equations, using the given initial conditions, to find

$$\begin{aligned} R(\omega) &= e^{H/2} R(0) + O(\omega^{-1/2}), \\ \varphi(\omega) &= \bar{\varphi}(\omega) + O(\omega^{-1/2}) \\ &= \varphi(0) + E(e^H - 1) \cdot \frac{H}{1 + (H/2\omega)^2} \frac{\alpha^2}{\omega} \csc^2(\varphi(0) - \psi) + O(\omega^{-1/2}). \end{aligned}$$

Undoing the substitutions then gives, for the solution of (2.8) subject to initial conditions (2.9),

$$\begin{aligned} r|_{x=1} &= e^{H/2} \frac{H}{\sqrt{1 + (H/2\omega)^2}} \frac{\alpha}{\omega} \csc(\theta(0) - \psi) + O(\omega^{-1/2}), \\ \theta|_{x=1} &= \theta(0) - \omega + \frac{EH(e^H - 1)}{1 + (H/2\omega)^2} \frac{\alpha^2}{\omega} \csc^2(\theta(0) - \psi) + O(\omega^{-1/2}). \end{aligned} \tag{2.13}$$

This approximation is valid as long as $r(0) = O(\omega^{-1/2})$. If we choose $\theta(0)$ to minimize the leading terms in this approximation to $\theta(1)$ we find that

$$\theta(0) = \psi + O\left[\left(\frac{\alpha^2}{\omega}\right)^{1/3}\right]$$

gives the approximation

$$\min \theta(1) = -\omega + O\left[\left(\frac{\alpha^2}{\omega}\right)^{1/3}\right] + O(\omega^{-1/2}).$$

For this choice of θ we have

$$r(0) = O\left(\frac{\alpha^{1/2}}{\omega^{2/3}}\right)$$

so that the approximation is valid.

This says that the trajectory, which starts on the line of initial conditions with $\theta(0)=\theta_{\min}$, winds $[\omega/2\pi]$ times clockwise around the spiral point. In the original coordinate system (y_1, y_2) this corresponds to a point η^* with the property asserted in the statement of Lemma 2.2.

3. The coupled equations. Now we show that the multiplicity results of § 2 hold for the full system (1.2)-(1.3). To begin, we note that if y is known then z is determined.

LEMMA 3.1. *Let $y \in C[0,1]$ be given. Then the boundary value problem for z, (1.3), has a unique solution.*

Proof. Write $q(x)=\gamma^{-1}\lambda \exp(h(y(x)))$. The difference of two solutions of (1.3) satisfies

$$\frac{1}{M}z''-z'-q(x)z=0$$

with the same (homogeneous) boundary conditions. Multiply this equation by $-z$ and integrate from zero to 1; integrate by parts and use the boundary conditions to get

$$0=\frac{1}{2}[z^2(1)+z^2(0)]+\frac{1}{M}\int_0^1 (z')^2\,dx+\int_0^1 qz^2\,dx.$$

Since $q>0$, this shows that $z=0$.

For $f\in C[0,1]$ write $\|f\|$ for $\max_{0\leqq x\leqq 1}|f(x)|$. We next bound $\|z\|$ in terms of bounds on y.

LEMMA 3.2. *Let Y be a positive constant. Then there are positive constants $\gamma_0=\gamma_0(\lambda, Y)$ and C such that for any $y\in C[0,1]$ with $0\leqq y(x)\leqq Y$ for $0\leqq x\leqq 1$, the solution z of (1.3) is bounded by*

$$\|z\|\leqq C\lambda\, e^{Y}$$

provided $\gamma\geqq\gamma_0$.

Proof. The solution z of

$$\begin{aligned}&\frac{1}{M}z''-z'=f(x), \qquad 0<x<1,\\ &z'(0)-Mz(0)=0,\\ &z'(1)=0\end{aligned}\tag{3.1}$$

is

$$z(x)=-\int_0^x f(\xi)\,d\xi-e^{Mx}\int_x^1 e^{-M\xi}f(\xi)\,d\xi.$$

A short calculation shows that $\|z\|\leqq\|f\|$. The solution z of (1.3) solves (3.1) with

$$f=\lambda \exp[h(y)](1+\gamma^{-1}z).$$

If $0\leqq y\leqq Y$ then $0\leqq h(y)<Y$, so that

$$\|z\|\leqq\lambda\, e^{Y}(1+\gamma^{-1}\|z\|).$$

Thus if $\gamma^{-1}\lambda\, e^{Y}<1$, then

$$\|z\|\leqq(1-\gamma^{-1}\lambda\, e^{Y})^{-1}\lambda\, e^{Y}$$

and the lemma follows.

Next we use this estimate to show that the solution of (1.3) depends Lipschitz continuously on the coefficient y, provided $0\leqq y\leqq Y$.

LEMMA 3.3. *If* $Y>0$, $y_1, y_2 \in C[0,1]$, *if* z_1 (*respectively*, z_2) *is the solution of* (1.3) *with* $y=y_1$ (*respectively*, $y=y_2$), *if* $0 \leqq y_k(x) \leqq Y$, $k=1,2$, *and if* $\gamma \geqq \gamma_0$, *then*

$$\|z_1-z_2\| \leqq (1-\gamma^{-1}\lambda\, e^Y)^{-1}(1+\gamma^{-1}C\lambda\, e^Y)\lambda\, e^Y \|y_1-y_2\|.$$

Proof. Subtract the equations for z_1 and z_2. Then

$$\frac{1}{M}(z_1-z_2)''-(z_1-z_2)' = \lambda(e^{h(y_1)}-e^{h(y_2)})+\gamma^{-1}\lambda(e^{h(y_1)}z_1-e^{h(y_2)}z_2)$$

$$= \lambda(e^{h(y_1)}-e^{h(y_2)})+\gamma^{-1}\lambda(e^{h(y_1)}-e^{h(y_2)})z_1+\gamma^{-1}\lambda\, e^{h(y_2)}(z_1-z_2),$$

and z_1-z_2 satisfies the boundary conditions. By Lemma 3.2, $\|z_1\| \leqq C\lambda e^Y$. The function $\exp[h(y)]$ is Lipschitz continuous in y for $0 \leqq y \leqq Y$ with Lipschitz constant e^Y. By the proof of Lemma 3.2,

$$\|z_1-z_2\| \leqq \lambda\, e^Y \|y_1-y_2\| + \gamma^{-1}\lambda\, e^Y \|y_1-y_2\| C\lambda\, e^Y + \gamma^{-1}\lambda\, e^Y \|z_1-z_2\|$$

and the lemma follows.

Next we estimate a Lipschitz constant for the dependence of the solution y of (1.2) on z regarded as a coefficient. It is nearly evident from the form of the equations that if y_1, y_2 are solutions of (1.2) with coefficients $z=z_1$ and $z=z_2$, respectively, and with the same initial data, then we have an estimate of the form

$$\|y_1-y_2\| \leqq \text{const}\ \gamma^{-1}\|z_1-z_2\|. \tag{3.2}$$

We could next combine this result with Lemma 3.3 to set up a convergent iterative scheme to solve the coupled boundary value problems (1.2)-(1.3). However, the constant—obtained in (3.2) from a direct application of the Gronwall inequality to (1.2)—would contain a term of the form $\exp(\exp\sqrt{\beta})$, and our iteration could be shown to converge only for enormously large γ. Therefore we localize (1.2) at the spiral point $(\alpha, 0)$ and obtain an estimate like (3.2) with a somewhat better constant.

Recall from Appendix C the change of variables localizing (1.4) at the spiral point

$$\begin{bmatrix} y-\alpha \\ y' \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ H/2 & \omega \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix}. \tag{3.3}$$

Apply the same change of variables to (1.2). If (y_k, y_k') corresponds to (u_k, v_k), $k=1,2$, under this change of variables—here subscripts indicate different solutions rather than components of a vector—then $u_1-u_2=y_1-y_2$ and we can estimate u_1-u_2 instead. The following proposition gives the resulting estimate.

PROPOSITION 3.1. *Consider the initial value problems* $k=1,2$:

$$\begin{aligned} &\frac{1}{H}y_k''-y_k'-\beta y_k+B\lambda(1+\gamma^{-1}z_k)\exp[h(y_k)]=0, \\ &y_k(0)=\eta, \\ &y_k'(0)=H\eta. \end{aligned} \tag{3.4}$$

Assume $0<\eta<\bar{\eta}$, $\|z_k\| \leqq C\lambda\, e^Y$ *for* $k=1,2$, *and that* γ *is sufficiently large. Then* $\|y_k\| < Y$ *for* $k=1,2$ *and there is a constant* c *such that*

$$\|y_1-y_2\| \leqq \gamma^{-1}Q\omega c \log\beta \exp[Q\omega\, e^{H/2}c\log\beta(1+\gamma^{-1}C\lambda\, e^Y)]\|z_1-z_2\|.$$

Use the change of variables (3.3) to transform (3.4) to

$$\begin{bmatrix} u_k' \\ v_k' \end{bmatrix} = A\begin{bmatrix} u_k \\ v_k \end{bmatrix} - Q\omega(N(u_k)+\gamma^{-1}z_kF(u_k))\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad k=1,2 \tag{3.5$_k$}$$

with appropriate initial conditions, as in Appendix C. Subtract (3.5_2) from (3.5_1) to obtain for the difference variables $(u, v) = (u_1 - u_2, v_1 - v_2)$:

$$\begin{aligned} \begin{bmatrix} u' \\ v' \end{bmatrix} &= A\begin{bmatrix} u \\ v \end{bmatrix} - Q\omega\{N(u_1) - N(u_2) + \gamma^{-1}[z_1F(u_1) - z_2F(u_2)]\}\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \\ \begin{bmatrix} u \\ v \end{bmatrix}(0) &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{aligned} \tag{3.6}$$

Now use the variation of parameters formula to get

$$\begin{bmatrix} u \\ v \end{bmatrix}(x) = -Q\omega \int_0^x e^{A(x-s)} W(s) \begin{bmatrix} 0 \\ 1 \end{bmatrix} ds,$$

in which W denotes the expression in curly brackets in (3.6). Now $u(x)$ is just the first component of the right-hand side, and we get

$$u(x) = -Q\omega \int_0^x e^{H(x-s)/2} \sin \omega(x-s)\, W(s)\, ds. \tag{3.7}$$

Write $N_k = N(u_k)$, $F_k = F(u_k)$; then

$$W = N_1 - N_2 + \gamma^{-1}[z_1(F_1 - F_2) + F_2(z_1 - z_2)]. \tag{3.8}$$

LEMMA 3.4. *There is a constant c such that $c \log \beta$ is a Lipschitz constant for N and F and a bound for F for $u \leqq Y - \alpha$.*

This follows from the forms of N and F derived in Appendix C, and from (B2) for Y. Next, by the bounds on y when $\gamma^{-1} = 0$ derived in § 2, the assumed bounds on η and $\|z_k\|$, $k = 1, 2$, and continuous dependence, it follows that $y_k < Y$ and thus $u_k < Y - \alpha$. Insert the Lipschitz conditions for N and F into (3.8) to obtain Lemma 3.5.

LEMMA 3.5. $|W(s)| \leqq c \log \beta(1 + \gamma^{-1} C\lambda\, e^Y)|u(s)| + \gamma^{-1} c \log \beta \|z_1 - z_2\|$, *and consequently*

$$|u(x)| \leqq Q\omega\, e^{H/2} \cdot c \log \beta(1 + \gamma^{-1} C\lambda\, e^Y) \int_0^x |u(s)|\, ds + \gamma^{-1} c \log \beta Q\omega\, e^{H/2} \|z_1 - z_2\|.$$

The estimate of Proposition 3.1 now follows by the Gronwall inequality [H].

PROPOSITION 3.2. *Let η be given, $0 < \eta < \bar{\eta}$. There is a unique ζ, $0 < \zeta < C\lambda\, e^Y$ such that the solution of the coupled equations* (1.2)-(1.3) *subject to the initial conditions*

$$\begin{aligned} y(0) &= \eta, \qquad y'(0) = H\eta, \\ z(0) &= \zeta, \qquad z'(0) = M\zeta, \end{aligned} \tag{3.9}$$

satisfies the second boundary condition for z: $z'(1) = 0$.

Proof. Construct sequences $\{y_k\}_{k=0}^{\infty}$ and $\{z_k\}_{k=0}^{\infty}$ as follows. Let $z_0 = 0$, and let y_0 be the solution of (1.2) with $z = z_0$ there, subject to the initial conditions (3.9). Then for each $k = 1, 2, \cdots$, let z_k be the (unique) solution of the boundary value problem (1.3) with $y = y_{k-1}$ there, and y_k the solution of the initial value problem (1.2), (3.9) with $z = z_k$. By Lemma 3.2 and Proposition 3.1 we have $\|z_k\| \leqq C\lambda\, e^Y$, $\|y_k\| \leqq Y$, $k \geqq 0$. Combine the estimates of Lemma 3.3 and Proposition 3.1 to get the estimate

$$\|y_{k+1} - y_k\| \leqq \gamma^{-1} C(\beta, H, B, \lambda) \|y_k - y_{k-1}\|.$$

A similar estimate holds for z. Thus if γ is sufficiently large, the y_k and z_k converge uniformly; the limiting pair (y, z) satisfy the differential equations (1.2)-(1.3), the initial conditions (3.9) for y, and the boundary conditions (1.3_0) for z.

COROLLARY. *The coupled boundary value problem* (1.2)-(1.3) *has as many solutions with* $0 \leqq y(0) \leqq \bar{\eta}$ *as has the decoupled equation* (1.4).

Proof. It is enough to show that (2.3) and the analogues of Lemmata 2.1 and 2.2 continue to hold when y is taken to be the solution of the coupled system. The angle $\tilde{\theta}$ is well defined: the polar radius r is bounded away from zero along solutions of the decoupled equation that start on the initial conditions. For sufficiently large γ this is still true for the solution of the full equation. Thus $\hat{\theta}$ is well defined, and the stated conclusions are valid by continuous dependence.

4. Comments.

(1) The estimate following (3.7) can probably be improved. The integral should be $O(\omega^{-1})$ unless W oscillates with frequency ω—but in that case the amplitude of W should be quite small.

(2) The leading term in the asymptotic expansion of θ, (2.13), is unimodular. If θ itself were indeed unimodular, then Lemma 2.1 would give an exact count of the number of solutions. Possibly, this could be answered by computing the solution of the variational equation by the method of averaging.

(3) The question of the stability of the steady states found here remains open.

Appendix A. Locate and classify critical points of the decoupled y-equation. The critical points of (2.1) are given by

$$\text{(A1)} \qquad y_2 = 0, \qquad y_1 = \frac{B\lambda}{\beta} \exp\left[\frac{y_1}{1+\gamma^{-1}y_1}\right].$$

Since γ is large, we expect the second equation to have two roots near the roots of

$$y = \frac{B\lambda}{\beta} e^y$$

when this has two roots, namely, when $B\lambda/\beta < 1/e$. In what follows we will choose successively large $\gamma = \gamma_0(\beta)$, with $\beta/\gamma_0(\beta) \to 0$ as $\beta \to \infty$; the expression $O(f(\beta))$ will mean "with a constant independent of γ, for all $\gamma > \gamma_0(\beta)$." For all results prior to § 3, it suffices to take $\gamma_0(\beta) = \beta^2$.

Iterate (A1) to obtain the first root

$$\text{(A2)} \qquad \alpha_0 = \frac{B\lambda}{\beta}\left(1 + \frac{B\lambda}{\beta} + O(\beta^{-2})\right).$$

Next, take logarithms in (A1):

$$\text{(A3)} \qquad \frac{y}{1+\gamma^{-1}y} = \log\frac{\beta}{B\lambda} + \log y.$$

Let $G = \beta/B\lambda$. Iterate the last equation, bearing in mind that γ is large; conclude that

$$\text{(A4)} \qquad \alpha \equiv \alpha_1 = \log G + \log\log G + \frac{\log\log G}{\log G} + O\left(\frac{\log\log\beta}{\log\beta}\right)^2.$$

This critical point takes a central role in the analysis that follows. We will always call it simply α.

The third critical point plays no role in this work, for the solutions that interest us will obey $0 \leqq y(x) \leqq Y$, with Y a β-dependent constant slightly larger than α. For

completeness, we give the third critical point. Rearrange (A3) to get

$$\log y = -\log G + \frac{y}{1+\gamma^{-1}y}$$

$$= -\log G + \frac{\gamma}{1+y^{-1}\gamma}.$$

We seek a root α_2 near $(B\lambda/\beta)\, e^{\gamma}$. Iterating now for $\log y$ gives

$$\alpha_2 = \frac{B\lambda}{\beta}\, e^{\gamma}(1+O(\gamma^2\, e^{-\gamma})).$$

Of course this is enormous compared to α.

Let us now classify the critical points. Denote the term in the exponential by

(A5) $$h(y) = \frac{y}{1+\gamma^{-1}y}.$$

Then a short calculation shows that the Jacobian of the system (2.1) at a critical point is given by

$$J(\alpha_j, 0) = \begin{bmatrix} 0 & 1 \\ H\beta(1-\alpha_j h'(\alpha_j)) & H \end{bmatrix}.$$

(This calculation uses the fact that (A1) holds for $y_1 = \alpha_j$.) The determinant is negative, and we have a saddle, for $(\alpha_0, 0)$ and $(\alpha_2, 0)$. The determinant is positive at $(\alpha, 0) = (\alpha_1, 0)$. At this point the discriminant $(\operatorname{Tr} J)^2 - 4 \det J$ is negative, and we have an unstable spiral, when the inequality

$$\alpha h'(\alpha) - 1 > \frac{H}{4\beta}$$

holds. Since $h'(y) = (1+\gamma^{-1}y)^{-2}$, this inequality holds as soon as β is large enough for any fixed H, B, λ; recall that α is given by (A4) and $\gamma^{-1}\beta$ is small.

Appendix B. The conservative system. Consider the potential energy function $V(y, \gamma^{-1})$ defined by

$$V'(y, \gamma^{-1}) = H\left(B\lambda \exp\left[\frac{y}{1+\gamma^{-1}y}\right] - \beta y\right),$$

$$V(0) = 0.$$

For the special case $\gamma^{-1} = 0$ we have explicitly

$$V(y, 0) = H\left(B\lambda(e^{y}-1) - \frac{\beta}{2}y^2\right).$$

The saddle-loop for the conservative system (2.4) crosses the y_1-axis at the point $(Y, 0)$ defined by

(B1) $$V(Y, \gamma^{-1}) = V(\alpha_0, \gamma^{-1}).$$

We determine Y for $\gamma^{-1} = 0$; then the Y for finite γ differs from this by $O(\gamma^{-1}\log\beta)$. We have to solve the following equation, from (B1) and the approximation (A2) for α_0:

$$H\left(B\lambda(e^{Y}-1) - \frac{\beta}{2}Y^2\right) = \frac{H}{2}\frac{(B\lambda)^2}{\beta} + O(\beta^{-2}).$$

This leads to

$$Y = \log\left(\frac{\beta Y^2}{2B\lambda} + 1 + O(\beta^{-1})\right)$$

$$= \log\frac{\beta}{B\lambda} + 2\log Y - \log 2 + O(\beta^{-1}),$$

and the solution is given by

$$Y = \log\frac{\beta}{B\lambda} + 2\log\log\frac{\beta}{B\lambda} - \log 2 + o(1)$$

$$= \alpha + \log\log\frac{\beta}{B\lambda} - \log 2 + o(1)$$

$$= \alpha + \log\log\sqrt{\beta/B\lambda} + o(1).$$

In addition, we find that the saddle-loop crosses the line of initial conditions at point η:

$$K(\eta, H\eta) = K(\alpha_0, 0).$$

The equation for the "large" solution leads to

$$\frac{H^2}{2}\eta^2 + H\left(B\lambda(e^\eta - 1) - \frac{\beta}{2}\eta^2\right) = O(\beta^{-1})$$

or

$$\eta = \log\left(\frac{\beta - H}{2B\lambda}\eta^2 + 1 + O(\beta^{-1})\right)$$

$$= \log\left(\frac{\beta}{B\lambda}\cdot\eta^2\cdot\frac{1 - H/\beta}{2} + 1 + O(\beta^{-1})\right)$$

and we find that $\eta = Y + O(\beta^{-1})$.

Appendix C. Localization at the spiral point. Translate the origin to the spiral point in (1.4) by the substitution

$$y_1 = w_1 + \alpha, \qquad y_2 = w_2$$

and get

$$\begin{pmatrix} w_1' \\ w_2' \end{pmatrix} = \begin{pmatrix} w_2 \\ H\{w_2 + \beta(w_1 + \alpha) - (B\lambda/\beta)\exp[h(\alpha + w_1)]\} \end{pmatrix}$$

$$= \begin{bmatrix} 0 & 1 \\ H\beta(1 - \alpha h'(\alpha)) & H \end{bmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

$$-H\alpha\beta\left(\exp\left[\frac{h'(\alpha)w_1}{1 + (\gamma^{-1}w_1)/(1 + \gamma^{-1}\alpha)}\right] - h'(\alpha)w_1 - 1\right)\begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The Jacobian at the spiral point is

$$J(\alpha) = \begin{bmatrix} 0 & 1 \\ H\beta(1 - \alpha h'(\alpha)) & H \end{bmatrix}$$

with complex eigenvalues given by

$$\frac{H}{2} \pm i\omega, \qquad \omega^2 = H\beta(\alpha h'(\alpha) - 1) - \frac{H^2}{4}.$$

Perform a linear change of coordinates with the matrix S,

$$S = \begin{bmatrix} 1 & 0 \\ H/2 & \omega \end{bmatrix},$$

whose columns are the real and imaginary parts of an eigenvector for $H/2 + i\omega$; substitute

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = S \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} u_1 \\ (H/2)u_1 + \omega u_2 \end{bmatrix};$$

and obtain the system

$$\begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \begin{bmatrix} H/2 & \omega \\ -\omega & H/2 \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} - Q\omega N(u_1) \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{C1}$$

Here we set

$$Q := \frac{H\alpha\beta}{\omega^2} = \frac{1}{h'(\alpha) - (1/\alpha)(1 + (H/4\beta))} = 1 + O(\alpha^{-1}),$$

$$N(u) := \exp\left[\frac{h'(\alpha)u}{1 + (\gamma^{-1}u)/(1 + \gamma^{-1}\alpha)}\right] - h'(\alpha)u - 1 \tag{C2}$$

is the nonlinearity. Note that when γ, β are large we have $Q \sim 1$; note further that $N(0) = N'(0) = 0$.

By following the coordinate changes in the boundary conditions we find in the new variables the line of initial conditions

$$u_2(0) = \frac{H}{2\omega} u_1(0) + \frac{H\alpha}{\omega}, \tag{C3}$$

and the line of final conditions

$$u_2(1) = -\frac{H}{2\omega} u_1(1). \tag{C4}$$

We can also perform the same change of variables for y in (1.2) with the coefficient z. We obtain

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}' = \begin{bmatrix} H/2 & \omega \\ -\omega & H/2 \end{bmatrix} \begin{pmatrix} u_1 \\ v_2 \end{pmatrix} - Q\omega(N(u_1) + \gamma^{-1} z F(u_1)) \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

with N as before and F given by

$$F(u) = \exp\left[\frac{h'(\alpha)u}{1 + (\gamma^{-1}u)/(1 + \gamma^{-1}\alpha)}\right].$$

Appendix D. Averaging computations. We begin by expanding the vector field of (2.10) in powers of $\omega^{-1/2}$. For this we require the Taylor expansion of the nonlinear

function N defined by (C2). Let

$$\sigma = \frac{\gamma^{-1}}{1+\gamma^{-1}\alpha}, \qquad k = h'(\alpha) = (1+\gamma^{-1}\alpha)^{-2},$$

the function h being defined by (A5). Then

$$N(u) = N_2u^2 + N_3u^3 + O(u^4) \tag{D1}$$

with

$$N_2 = \frac{k^2}{2} - \sigma k, \qquad N_3 = \frac{k^3}{6} - \sigma k^2 + \sigma^2 k.$$

We note that $N_2 = \frac{1}{2} + O(\gamma^{-1}\alpha)$, $N_3 = \frac{1}{6} + O(\gamma^{-1}\alpha)$. This illustrates the earlier remark that replacing $h(y)$ by y in the exponential throughout makes no essential difference.

Inserting the expansion (D1) for N into (2.10) leads to

$$\begin{aligned} \dot{R} &= -\omega^{-1/2}QR^2N_2\cos^2(\varphi - t)\sin(\varphi - t) \\ &\quad + \omega^{-1}\left(\frac{H}{2}R - QR^3N_3\cos^3(\varphi - t)\sin(\varphi - t)\right) + O(\omega^{-3/2}), \\ \dot{\varphi} &= -\omega^{-1/2}QRN_2\cos^3(\varphi - t) - \omega^{-1}QR^2N_3\cos^4(\varphi - t) + O(\omega^{-3/2}). \end{aligned} \tag{D2}$$

Define

$$f(t, R, \varphi) = -QN_2\begin{pmatrix} R^2\cos^2(\varphi - t)\sin(\varphi - t) \\ R\cos^3(\varphi - t) \end{pmatrix},$$

$$g(t, R, \varphi) = \begin{pmatrix} (H/2)R - QR^3N_3\cos^3(\varphi - t)\sin(\varphi - t) \\ -QR^2N_3\cos^4(\varphi - t) \end{pmatrix}.$$

Then (D2) may be written

$$(\dot{R}, \dot{\varphi})^T = \omega^{-1/2}f(t, R, \varphi) + \omega^{-1}g(t, R, \varphi) + \omega^{-3/2}p(t, R, \varphi, \omega^{-1/2})$$

with $\omega^{-3/2}p$ being the remainder in Taylor's formula. The functions f, g, p are 2π-periodic in t and the average of f vanishes:

$$f^0(R, \varphi) = \frac{1}{2\pi}\int_0^{2\pi} f(\tau, R, \varphi)\, d\tau = 0.$$

Define

$$v^1(t, R, \varphi) = \int_0^t f(\tau, R, \varphi)\, d\tau - \frac{1}{2\pi}\int_0^{2\pi}\int_0^t f(\tau, R, \varphi)\, d\tau\, dt,$$

$$f^1(t, R, \varphi) = D_{(R,\varphi)}f(t, R, \varphi)v^1(t, R, \varphi)$$

and let f^{10} and g^0 be the respective mean values over a period of f^1 and g. By Theorem 3.9.1 of [SV], the solution of (2.10) with initial conditions $R(0) = R_0$, $\varphi(0) = \varphi_0$ is approximated with an error that is $O(\omega^{-1/2})$ for $0 \leqq t \leqq \omega$ by the solution of

$$\begin{aligned} &\frac{d}{dt}(\bar{R}, \bar{\varphi})^T = \omega^{-1}(f^{10}(\bar{R}, \bar{\varphi}) + g^0(\bar{R}, \bar{\varphi})), \\ &(\bar{R}(0), \bar{\varphi}(0)) = (R_0, \varphi_0). \end{aligned} \tag{D3}$$

We now compute v^1, f^1, f^{10}, and g^0:

$$v^1(t, R, \varphi) = -QRN_2 \begin{pmatrix} \int_0^t R\cos^2(\varphi-\tau)\sin(\varphi-\tau)\,d\tau - \text{Mean value} \\ \int_0^t \cos^3(\varphi-\tau)\,d\tau - \text{Mean value} \end{pmatrix}$$

$$= -QRN_2 \begin{pmatrix} \frac{1}{3}R\cos^3(\varphi-\tau)_0^t - \text{Mean value} \\ -\sin(\varphi-\tau)_0^t + \frac{1}{3}\sin^3(\varphi-\tau)_0^t - \text{Mean value} \end{pmatrix}$$

$$= -QRN_2 \begin{pmatrix} \frac{1}{3}R\cos^3(\varphi-t) - \frac{1}{3}R\cos^3\varphi - \text{Mean value} \\ -\sin(\varphi-t) + \sin\varphi + \frac{1}{3}\sin^3(\varphi-t) - \frac{1}{3}\sin^3\varphi - \text{Mean value} \end{pmatrix}$$

$$= -QRN_2 \begin{pmatrix} \frac{R}{3}\cos^3(\varphi-t) \\ -\sin(\varphi-t) + \frac{1}{3}\sin^3(\varphi-t) \end{pmatrix}.$$

Next,

$$f^1 = D_{(R,\varphi)} f \cdot v^1$$

$$= Q^2N_2^2R \begin{pmatrix} 2R\cos^2(\varphi-t)\sin(\varphi-t) & R^2[\cos(\varphi-t) - 3\sin^2(\varphi-t)\cos(\varphi-t)] \\ \cos^3(\varphi-t) & -3R\cos^2(\varphi-t)\sin(\varphi-t) \end{pmatrix}$$

$$\cdot \begin{pmatrix} \frac{R}{3}\cos^3(\varphi-t) \\ -\sin(\varphi-t) + \frac{1}{3}\sin^3(\varphi-t) \end{pmatrix}$$

$$= Q^2N_2^2R$$

$$\cdot \begin{pmatrix} R^2\left(\frac{2}{3}\cos^5(\varphi-t)\sin(\varphi-t) + \cos(\varphi-t)(1-3\sin^2(\varphi-t))\sin(\varphi-t)\left(\frac{1}{3}\sin^2(\varphi-t) - 1\right)\right) \\ R\left(\frac{1}{3}\cos^6(\varphi-t) + 3\cos^2(\varphi-t)\sin^2(\varphi-t) - \cos^2(\varphi-t)\sin^4(\varphi-t)\right) \end{pmatrix}.$$

To compute f^{10}, observe that the first component of f^1 has mean value zero. Hence

$$f^{10}(R, \varphi) = \frac{5}{12}Q^2N_2^2R^2\begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Next,

$$g^0(R, \varphi) = \begin{pmatrix} \frac{H}{2}R \\ -\frac{3}{8}QN_3R^2 \end{pmatrix}.$$

So the averaged equations are

$$\frac{d\bar{R}}{dt} = \omega^{-1}\frac{H}{2}\bar{R}, \qquad \bar{R}(0) = R_0,$$

$$\frac{d\bar{\varphi}}{dt} = \omega^{-1}E\bar{R}^2, \qquad \bar{\varphi}(0) = \varphi_0, \tag{D4}$$

$$E = \frac{5}{12}Q^2N_2^2 - \frac{3}{8}QN_3 = \frac{1}{24} + O(\alpha^{-1}).$$

Acknowledgments. It is a pleasure to thank Jim Murdock for useful discussions, and H. F. Weinberger for his hospitality at the Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, Minnesota.

REFERENCES

[A1] R. ALEXANDER, *The multiple steady states of a nonadiabatic tubular reactor*, J. Math. Anal. Appl., 101 (1984), pp. 12-22.

[A2] ———, *Mathematical analysis of tubular reactors*, in Reacting Flows: Combustion and Chemical Reactors, G. S. S. Ludford, ed., American Mathematical Society, Providence, RI, 1986, Vol. 2, pp. 317-330.

[H] P. HARTMAN, *Ordinary Differential Equations*, Hartman, Baltimore, MD, 1973.

[KP] A. K. KAPILA AND A. B. POORE, *The steady response of a nonadiabatic tubular reactor: new multiplicities*, Chem. Engrg. Sci., 37 (1982), pp. 57-76.

[MA] L. MARKUS AND N. AMUNDSON, *Nonlinear boundary value problems arising in chemical reactor theory*, J. Differential Equations, 4 (1968), pp. 102-113.

[SV] J. A. SANDERS AND F. VERHULST, *Averaging Methods in Nonlinear Dynamical Systems*, Springer-Verlag, Berlin, New York, 1985.

[VA] A. VARMA AND R. ARIS, *Stirred pots and empty tubes*, in Chemical Reactor Theory: A Review, L. Lapidus and N. R. Amundson, eds., Prentice-Hall, Englewood Cliffs, NJ, 1977, pp. 79-155.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 154-178, January 1990
© 1990 Society for Industrial and Applied Mathematics
009

INERTIAL MANIFOLDS AND MULTIGRID METHODS*

R. TEMAM†

Abstract. This article presents an analogy existing between the concepts of approximate inertial manifolds in dynamical systems theory and multigrid methods in numerical analysis. In view of the large-time approximation of dissipative evolution equations in a turbulent regime, a new algorithm is proposed and studied that combines some ideas and concepts of inertial manifolds and multigrid methods. This article emphasizes theoretical questions. More practical (computational) questions will be investigated elsewhere.

Key words. inertial manifolds, attractors, partial differential equations, approximation, multigrid methods

AMS(MOS) subject classifications. 35K60, 65N05

Introduction. Two theories have developed in parallel during the last years with different objectives; namely, the theory of inertial manifolds that has emerged from the study of dynamical systems and the theory of multigrid methods in numerical analysis. Although these two theories seem very far apart, our aim here is to show that they have some underlying ideas in common and to investigate the relation between them.

Multigrid methods concern the numerical solution of partial differential equations by finite differences or finite elements using two (or more) mesh grids, one finer and one coarser. The main observation constituting the starting point of the theory is that a simple iterative method is sufficient to determine the high-frequency components of the solution, but that further effort is needed to solve the low-frequency components (see [B], [H], [Mc], and the references therein).

In dynamical systems theory the objective is to study the long-term behavior of the solutions of an evolution equation. When the equation is dissipative all solutions converge as $t \to \infty$ to a complicated set $\mathscr{A}$, the global attractor, which may be fractal. This set embodies the large-time dynamics of the equation, corresponding to all sorts of regimes, including the turbulent ones. Although this set may be fairly complicated, in general it has finite dimension. Inertial manifolds are smooth finite-dimensional manifolds that are invariant by the flow, contain the global attractor, and attract all the orbits at an exponential rate. All dissipative systems are not known to possess an inertial manifold, but the related concept of approximate inertial manifolds has been introduced and applies to a broad class of dissipative systems (see [FST1], [FST2], [FNST1], [FNST2], [CFNT], [MS], [T4], and the references therein). Now, in essence, inertial and approximate inertial manifolds correspond to an exact (or approximate) interaction law between small and large wavelengths. When an orbit lies on the inertial manifold the small wavelengths are, at each instant of time, an explicit function of the large wavelengths, this correlating function being the equation of the manifold. And of course, an orbit starting outside the manifold converges to it exponentially fast, and thus soon, after an initial period, the interaction law between small and large wavelengths goes into effect.

In previous articles (see [FMT], [T5], [MT1], [MT2], [JRT]), we have shown how we can construct effective numerical algorithms by using approximate inertial manifolds, in the context of spectral and finite-element methods. Our object in this

* Received by the editors January 31, 1989; accepted for publication April 6, 1989. This work was supported in part by National Science Foundation grant DMS-880296.

† Laboratoire d'Analyse Numérique, Bâtiment 425, Université Paris-Sud, 91405 Orsay, France, and Institute for Applied Mathematics and Scientific Computing, Indiana University, Bloomington, Indiana 47405.

article is to present and study similar algorithms in the case of finite differences. Here we naturally meet the methodology of multigrid methods: the large wavelength components of the flow are based on the coarse grid, while the small wavelength components are based on the fine grid. Let us mention, however, that despite this analogy, the questions that we address here are quite different from those addressed in the multigrid literature. The multigrid literature emphasizes more the linear part of the equation, while here we are more interested in the nonlinear terms: following ideas derived from inertial manifolds and dynamical systems theory, our object here is to describe in an approximate way the nonlinear interaction of small wavelengths based on the finer grid and large wavelengths based on the coarser grid.

For the sake of simplicity we will devote this article to some specific situations and examples, but our results apply to much more general situations. We consider a class of nonlinear evolution equations for which the linear elliptic part corresponds to a Dirichlet problem in a square. In § 1 we describe our space discretization procedure based on the use of the *incremental unknowns*. We are given two grids in the square, the fine and coarse grids; the discretization is simply the five-point discretization of the Laplace operator on the fine grid. However, instead of considering the usual nodal values for the unknown function, we will consider the incremental unknowns on the fine grid. They consist of the nodal values on the coarse grid and, on the points of the fine grids not belonging to the coarse one, the unknown is the increment to an interpolate value of the neighboring coarse points. We are not aware of any explicit use of the incremental unknowns in numerical analysis.[1] The material presented in § 1 is, however, very simple and necessary for the rest of the article.

In § 2 we present, at the level of linear elliptic problems, the appropriate variational setting for the use of the incremental unknowns. The variational framework for finite-difference discretization of linear elliptic problems was introduced by Céa [C]. Although the variational framework is very well suited for finite elements, and its use is routine in this case, it is not indispensable and is less used in the context of finite differences. We will see that it is very appropriate here, at least for the theoretical part of the work; we will recall and present the necessary material in § 2.

In § 3 we present the nonlinear equations that we study. An abstract equation and three specific equations related to the Navier–Stokes equations and to reaction-diffusion equations are presented. As previously mentioned, our results apply to general equations, but we refrain here from considering general equations and concentrate on the specific examples. In § 4 we implement the spatial discretization of the problem using two different grids and the incremental unknowns, and implement the algorithm based on inertial manifolds and multigrid methods (the IMG algorithm). The stability, consistency, and convergence of the algorithm are investigated in § 5, which relies extensively on the use of energy methods. Finally, in § 6 we consider a full discretized version of the IMG algorithm, i.e., one involving space and time discretization. We consider an explicit time discretization scheme and restrict ourselves to a *linear problem*. Indeed, another advantage of the IMG algorithm appears already, at the level of a linear evolution equation, in the form of an improved stability condition; namely, the stability condition is that corresponding to the coarse mesh instead of that corresponding to the fine mesh.

In this article we have emphasized simple equations and theoretical aspects in connection with dynamical systems. In subsequent works we intend to consider more general equations; and we intend to describe in more detail and in a more practical

[1] They appear in a hidden way in the multigrid methods through the restriction and prolongation operators.

form the computational aspects of the algorithm. Other related forms of the algorithm will also be considered; see in particular [T8].

1. Matricial structure of the problem. In this section we introduce the incremental unknowns and show their use in the solution of finite-difference problems. We consider successively one- and two-dimensional problems.

1.1. The one-dimensional case. For the sake of simplicity we start with the model one-dimensional problem:

$$-\frac{d^2u}{dx^2}=f \quad \text{in } (0,1), \qquad u(0)=u(1)=0. \tag{1.1}$$

For $N\in\mathbb{N}$, we consider the fine grid corresponding to the discretization mesh $h=1/2N$ and the coarse grid corresponding to the discretization mesh $2h=1/N$. The coarse gridpoints are the points $2jh$, $j=0,\cdots,N$ ($j=0$ and N correspond to the boundary points); the fine gridpoints are the points jh, $j=0,\cdots,2N$ (see Fig. 1.1). We write $f_j=f(jh)$, $u_j\simeq u(jh)$ and, on the fine grid, we write the usual finite difference scheme

$$-\frac{1}{h^2}(u_{j+1}-2u_j+u_{j-1})=f_j, \quad j=1,\cdots,2N-1, \quad u_0=u_{2N}=0. \tag{1.2}$$

The incremental unknowns consist of the numbers u_{2j}, $j=1,\cdots,N-1$, corresponding to the approximate values at the points $2jh$, and of the numbers

$$\bar{u}_{2j+1}=u_{2j+1}-\tfrac{1}{2}(u_{2j}+u_{2j+2}), \qquad j=0,\cdots,N-1, \tag{1.3}$$

corresponding to the increments from the average values at the neighbors, at the points $(2j+1)h$, $j=0,\cdots,N-1$. We easily infer from (1.2) that

$$\bar{u}_{2j+1}=\frac{h^2}{2}f_{2j+1}, \qquad j=0,\cdots,N-1, \tag{1.4}$$

$$\frac{1}{4h^2}\{2u_{2j}-u_{2j-2}-u_{2j+2}\}-\frac{1}{2}\{\bar{u}_{2j-1}+\bar{u}_{2j+1}\}=\frac{1}{2}f_{2j}, \qquad j=1,\cdots,N-1. \tag{1.5}$$

If we take (1.4) into account, (1.5) becomes

$$\frac{1}{4h^2}\{2u_{2j}-u_{2j-2}-u_{2j+2}\}=\frac{1}{2}f_{2j}+\frac{1}{4}\{f_{2j-1}+f_{2j+1}\}, \qquad j=1,\cdots,N-1. \tag{1.6}$$

The key point for the nonlinear analysis hereafter is that the incremental values are small. This is transparent in (1.4), which shows that

$$\bar{u}_{2j+1}=O(h^2). \tag{1.7}$$

In matricial form, we write (1.2) as

$$\tilde{A}\tilde{U}=\tilde{b}, \tag{1.8}$$

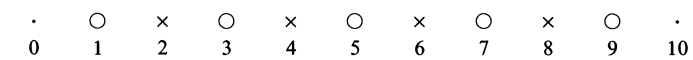


FIG. 1.1. *Coarse gridpoints* (○) *and fine gridpoints* (×, ○) *on* $(0,1)$ *for* $N=5$.

where $\tilde{U} = (u_1, \cdots, u_{2N-1})^t$, $\tilde{b} = (f_1, \cdots, f_{2N-1})^t$, and $\tilde{A} = (1/h^2)T$, T being the usual tridiagonal matrix

$$T = \begin{pmatrix} 2 & -1 & & & 0 \\ -1 & \ddots & \ddots & & \\ & \ddots & \ddots & & -1 \\ 0 & & & -1 & 2 \end{pmatrix}.$$

We may reorder $\tilde{U}$, $\tilde{b}$ into their coarse and fine components:

$$U = (U_c, U_f)^t, \qquad b = (b_c, b_f)^t,$$

$$U_c = (u_2, \cdots, u_{2N-2})^t, \qquad U_f = (u_1, u_3, \cdots, u_{2N-1})^t,$$

$$b_c = (f_2, \cdots, f_{2N-2})^t, \qquad b_f = (f_1, f_3, \cdots, f_{2N-1})^t,$$

and rewrite (1.8) as

$$AU = b, \tag{1.9}$$

this time with

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

$A_{11} = (2/h^2)I_{N-1}$, $A_{22} = (2/h^2)I_N$, where I_j is the jth-dimensional unit matrix and A_{12} is an $(N-1) \times N$ matrix:

$$A_{12} = A_{21} = -\frac{1}{h^2} B_N,$$

$$B_N = \begin{pmatrix} 1 & 1 & & & \\ & \ddots & \ddots & & 0 \\ & 0 & & & \\ & & & 1 & 1 \end{pmatrix}.$$

Using the incremental unknowns, we then replace U by $\hat{U} = (U_c, U_i)^t$,

$$U = S\hat{U}, \tag{1.10}$$

$$S = \begin{pmatrix} I_{N-1} & 0 \\ \frac{1}{2}B_N^t & I_N \end{pmatrix}, \qquad S^{-1} = \begin{pmatrix} I_{N-1} & 0 \\ -\frac{1}{2}B_N^t & I_N \end{pmatrix}.$$

We infer from (1.9) that $AS\hat{U} = b$ or

$$\hat{A}\hat{U} = \hat{b}, \tag{1.11}$$

with $\hat{A} = S^tAS$, $\hat{b} = S^tb$. Like A, the matrix $\hat{A}$ is symmetric positive definite.

1.2. The two-dimensional case. Although some steps of the procedure will now be less transparent, we will proceed in exactly the same way for the two-dimensional case. We consider the Dirichlet problem

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega \tag{1.12}$$

in the square $\Omega = (0, 1)^2$ and its five-point discretization. As above, for $N \in \mathbb{N}$, we set $h = 1/2N$, $2h = 1/N$ and consider the fine grid corresponding to the mesh h, and the

coarse grid corresponding to the mesh $2h$ in both directions. For $i, j = 0, \cdots, 2N$, we write $f_{ij} = f(ih, jh)$ and $u_{ij} \simeq u(ih, jh)$ is the approximate value of $u(ih, jh)$. The fine gridpoints are the points (ih, jh), $i = 1, \cdots, 2j-1$, $j = 1, \cdots, 2j-1$, and the coarse gridpoints are the points $(2ih, 2jh)$, $i = 1, \cdots, N-1$, $j = 1, \cdots, N-1$.

The discrete equations read

$$\frac{1}{h^2}(2u_{i,j} - u_{i-1,j} - u_{i+1,j}) + \frac{1}{h^2}(2u_{i,j} - u_{i,j-1} - u_{i,j+1}) = f_{ij}, \tag{1.13}$$

$$1 \leqq i, \qquad j \leqq 2N-1.$$

Usually we number the unknowns u_{ij} in a sequential order, say from left to right and from top to bottom, and reinterpret (1.13) as a system similar to (1.8). We may also, as for (1.9), reorder the nodes in a different way, with the coarse gridpoints first[2] (numbered from left to right and from top to bottom), then the rest of the fine gridpoints numbered in the same way.[3] We write

$$U = (U_c, U_f)^t, \qquad b = (b_c, b_f)^t,$$

and we then have an analogue of (1.9):

$$AU = b, \tag{1.14}$$

with a matrix A that will not be made explicit here.

At this point we introduce the incremental unknowns: they are made first of the coarse grid nodal values $u_{2i,2j}$, $i, j = 1, \cdots, N-1$. Then, at the noncoarse gridpoints, the incremental unknowns are defined as follows (see Fig. 1.3):

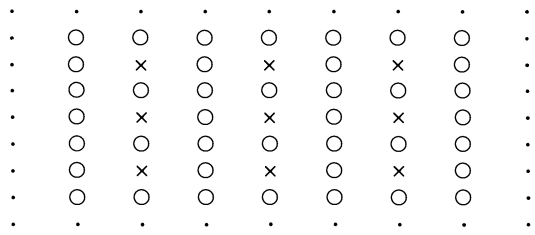


FIG. 1.2. *Coarse gridpoints* (×) *and fine gridpoints* (×, ○) *on the square* $(0, 1)^2$ *for* $N = 4$.

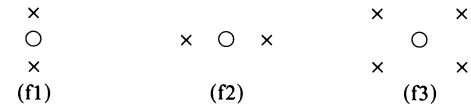


FIG. 1.3. *Noncoarse gridpoints of type* (f1), (f2), *or* (f3).

[2] See points × in Fig. 1.2.

[3] See points ○ in Fig. 1.2. In fact, a more subtle numbering of the noncoarse gridpoints (×) is desirable; this question will be addressed elsewhere.

—Fine gridpoints of type (f1), those at the middle of two vertical coarse gridpoints. The incremental unknown is then

$$\bar{u}_{2i+1,2j} = u_{2i+1,j} - \tfrac{1}{2}(u_{2i,j} + u_{2i+2,j}), \qquad i = 0, \cdots, N-1, \quad j = 1, \cdots, N-1. \tag{1.15}$$

—Fine gridpoints of type (f2), those at the middle of two horizontal coarse gridpoints. The incremental unknown at such a point reads

$$\bar{u}_{2i,2j+1} = u_{2i,2j+1} - \tfrac{1}{2}(u_{2i,2j} + u_{2i,2j+2}), \qquad i = 1, \cdots, N-1, \quad j = 0, \cdots, N-1. \tag{1.16}$$

—Fine gridpoints of type (f3), the rest of the noncoarse gridpoints. They are at the center of a square of edge $2h$, the vertices of which are coarse gridpoints (or boundary points). In this case we introduce the incremental unknown

$$\bar{u}_{2i+1,2j+1} = u_{2i+1,2j+1} - \tfrac{1}{4}(u_{2i,2j} + u_{2i,2j+2} + u_{2i+2,2j} + u_{2i+2,2j+2}), \qquad i, j = 0, \cdots, N-1. \tag{1.17}$$

Note that in (1.15)-(1.17), $u_{\alpha\beta} = 0$ if α or $\beta = 0$ or $2N$; of course the corresponding unknowns disappear then.

Let $\bar{U}_f$ denote the incremental unknowns defined by (1.15)-(1.17) and let $\hat{U} = (U_c, \bar{U}_f)$ be the new unknowns. We have

$$U = S\hat{U}, \tag{1.18}$$

where the matrix S, as well as its inverse, is easily derived from (1.15)-(1.17). Then (1.14) yields

$$\hat{A}\hat{U} = \hat{b} \tag{1.19}$$

with $\hat{A} = S^t A S$ and $\hat{b} = S^t b$.

The explicit form of $\hat{A}$ is less transparent here than in dimension 1. It is also less transparent that the incremental values $\bar{U}_f$ are small as in (1.4); however, this will be proved in § 2 by using the variational approach.

2. Variational framework. The variational formulation of (1.11) is well known. We introduce the Sobolev space $V = H_0^1(\Omega)$ endowed with its scalar product

$$a(u, v) = ((u, v)) = \int_\Omega \text{grad } u.\text{grad } v \, dx$$

and we look for $u \in V$ such that

$$a(u, v) = (f, v) \quad \forall v \in V, \tag{2.1}$$

where (f, v) is the $L^2(\Omega)$-scalar product of f and v:

$$(f, v) = \int_\Omega fv \, dx.$$

We denote by $\|\cdot\|$ and $|\cdot|$ the V and L^2 norms corresponding to the scalar products $((\cdot, \cdot))$, $(\cdot, \cdot)$, respectively.

We now recall briefly the variational framework corresponding to the finite-difference scheme (1.12) (see [C]). For the mesh $h = 1/2N$, we introduce the space V_h, which consists of step functions $u_h, v_h, \cdots$, that are constants on the squares centered at $(ih, (i+1)h) \times (jh, (j+1)h)$, $i, j = 0, \cdots, 2N-1$ of edge h and that vanish

if i or $j = 0$ or $2N-1$. The space V_h is spanned by the basis functions w_{hM}, $M = (ih, jh)$, $i, j = 1, \cdots, 2N-2$, which are equal to 1 in the square $[ih, (i+1)h) \times [jh, (j+1)h)$ and which vanish outside this square. Thus

$$u_h(x) = \sum_{M \in \mathring{\Omega}_h} u_h(M) w_{hM}(x), \qquad x \in \Omega; \tag{2.2}$$

$\mathring{\Omega}_h$ is the set of points (ih, jh), $i, j = 1, \cdots, 2N-2$, and we denote by Ω_h the set of points (ij, jh), $i, j = 0, \cdots, 2N$.

We introduce the finite-difference operators ∇_{1h}, ∇_{2h}:

$$\nabla_{ih}\varphi(x) = \frac{1}{h}(\varphi(x + he_i) - \varphi(x)),$$

$e_1 = (1, 0)$, $e_2 = (0, 1)$ and we endow V_h with the scalar product

$$((u_h, v_h))_h = \sum_{i=1}^{2} (\nabla_{ih} u_h, \nabla_{ih} v_h),$$

where $(\cdot, \cdot)$ is as before the scalar product in $L^2(\Omega)$. We set $\|\cdot\|_h = \{((\cdot, \cdot))_h\}^{1/2}$ and observe that $\|\cdot\|_h$ and $|\cdot|$ are Hilbert norms on V_h. The discrete analogue of (2.1) is the following variational problem:

(2.3) Find $u_h \in V_h$ such that $((u_h, v_h))_h = (f, v_h)$, for all $v_h \in V_h$.

Setting $u_h(M) = u_{ij}$ for $M = (ih, jh) \in \mathring{\Omega}_h$, it can easily be shown that (2.3) is equivalent to the system of equations (1.12) with the only difference that here f_{ij} is an average value of f:

$$f_{ij} = \frac{1}{h^2} \int_{ih}^{(i+1)h} \int_{jh}^{(j+1)h} f(x)\, dx. \tag{2.4}$$

Now there is no objection to considering the mesh $2h = 1/N$ and the corresponding space V_{2h} spanned by the basis functions $w_{2h,M}$, $M \in \mathring{\Omega}_{2h}$, where $w_{2h,M}$ and $\mathring{\Omega}_{2h}$ are defined exactly as before. We are not interested here in the analogue of (2.3) in V_{2h}. Rather we are interested in an appropriate rewriting of (2.3). We observe that $V_{2h} \subset V_h$ and write

$$V_h = V_{2h} \oplus W_h, \tag{2.5}$$

where W_h is the space spanned by the functions w_{hM}, $M \in \mathring{\Omega}_h \backslash \mathring{\Omega}_{2h}$. We thus obtain a basis of V_h consisting of the $w_{2h,M}$, $M \in \mathring{\Omega}_{2h}$, and the w_{hM}, $M \in \mathring{\Omega}_h \backslash \mathring{\Omega}_{2h}$. The previously used basis of V_h or V_{2h} is called the natural basis, while this basis of V_h induced by that of V_{2h} will be called the *induced basis.*

We now write the decomposition of an element $u_h \in V_h$ corresponding to (2.5):

$$u_h = y_h + z_h, \quad y_h \in V_{2h}, \quad z_h \in W_h, \tag{2.6}$$

and try to identify y_h and z_h, on a square $[2ih, 2(i+1)h[\times [2jh, 2(j+1)h[$ of vertex $M_1 = (2ih, 2jh)$, as in Fig. 2.1.

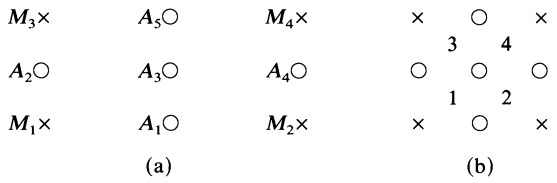


FIG. 2.1. *Subdivision of a coarse grid square into four fine grid squares.*

We have

$$w_{2h,M_1} = w_{hM_1} + \sum_{i=1}^{3} w_{hA_i}.$$

Hence

$$u_h = \sum_{M \in \mathring{\Omega}_h} u_h(M) w_{hM} \tag{2.7}$$

is equal on this square to

$$u_h(M_1) w_{hM_1} + \sum_{i=1}^{3} u_h(A_i) w_{hA_i} = u_h(M_1) w_{2hM_1} + \sum_{i=1}^{3} (u_h(A_i) - u_h(M_1)) w_{hA_i}.$$

We conclude that y_h is the coarse grid component

$$y_h = \sum_{M \in \mathring{\Omega}_{2h}} u_h(M) w_{2hM}, \tag{2.8}$$

and

$$\begin{aligned} z_h = \sum_{M \in \Omega_h} \{ & (u_h(M + he_1) - u_h(M)) w_{h,M+he_1} \\ & + (u_h(M + he_2) - u_h(M)) w_{h,M+he_2} \\ & + (u_h(M + he_1 + he_2) - u_h(M)) w_{h\,M+he_1+he_2} \}. \end{aligned} \tag{2.9}$$

The components of z_h in the basis of W_h described before are thus the incremental quantities of the form

$$u_h(A_1) - u_h(M_1), \quad u_h(A_2) - u_h(M_1), \quad u_h(A_3) - u_h(M_1). \tag{2.10}$$

We will then call z_h the *incremental component* of u_h and y_h its *coarse grid component.* Let us observe that

$$z_h(M) = 0 \quad \forall M \in \Omega_{2h}, \quad \forall z_h \in W_h. \tag{2.11}$$

Remark 2.1. These incremental values of u_h are not those used in § 1. A different basis of V_h leading to cumbersome computations must be used to recover the incremental unknowns mentioned in § 1. For the sake of simplicity we will pursue the analysis with incremental values (2.10) and decompositions (2.5)-(2.6), (2.8)-(2.9). However, for all practical purposes, we advocate the use of the incremental unknowns of § 1, which are much more convenient in effective computations.

We now return to the approximate problem, namely (2.3). Using the decomposition (2.5), (2.6) we write $u_h = y_h + z_h$ and observe that (2.3) is equivalent to

$$\begin{aligned} ((y_h + z_h, \tilde{y}_h))_h &= (f, \tilde{y}_h) \quad \forall \tilde{y}_h \in V_{2h}, \\ ((y_h + z_h, \tilde{z}_h))_h &= (f, \tilde{z}_h) \quad \forall \tilde{z}_h \in W_h. \end{aligned} \tag{2.12}$$

Of course, replacing v_h by u_h in (2.3) we obtain the usual a priori estimates:

$$\|u_h\|_h^2 = (f, u_h). \tag{2.13}$$

Thanks to the discrete Poincaré inequality (see [C] or [T2]),

$$|u_h| \leqq c_1 \|u_h\|_h, \tag{2.14}$$

where c_1 is independent of h, we find

$$\begin{aligned} \|u_h\|_h^2 &\leqq |f|\,|u_h| \leqq c_1 |f|\,\|u_h\|_h, \\ \|u_h\|_h &\leqq c_1 |f|. \end{aligned} \tag{2.15}$$

We can also obtain (2.13) by replacing $\tilde{y}_h$ by y_h in the first equation of (2.12) and $\tilde{z}_h$ by z_h in the second equation of (2.12), and adding the resulting relations.

Our aim is now to show that (2.15) in fact yields two separate a priori estimates for y_h and z_h.

Let us admit temporarily the following lemmas, which will be proved below.

LEMMA 2.1 (Strong Cauchy-Schwarz inequality). *We have the enhanced Cauchy-Schwarz inequality*:

$$|((y_h, z_h))_h| \leqq \frac{\sqrt{3}}{2} \|y_h\|_h \|z_h\|_h \quad \forall y_h \in V_{2h}, \quad \forall z_h \in W_h. \tag{2.16}$$

LEMMA 2.2.

$$\|y_h\|_h^2 = 2\|y_h\|_{2h}^2 \quad \forall y_h \in V_{2h}. \tag{2.17}$$

LEMMA 2.3 (Strong Poincaré inequality in W_h). *We have the following strong Poincaré inequality for functions in* W_h:

$$|z_h| \leqq h\|z_h\|_h \quad \forall z_h \in W_h. \tag{2.18}$$

If we admit these lemmas, then

$$\begin{aligned} \|u_h\|_h^2 &= \|y_h + z_h\|_h^2 \\ &= \|y_h\|_h^2 + \|z_h\|_h^2 + 2((y_h, z_h))_h \\ &\geqq \|y_h\|_h^2 + \|z_h\|_h^2 - \sqrt{3}\|y_h\|_h \|z_h\|_h \\ &\geqq \left(1 - \frac{\sqrt{3}}{2}\right)\{\|y_h\|_h^2 + \|z_h\|_h^2\} \\ &\geqq \frac{1}{8}\{\|y_h\|_{2h}^2 + \|z_h\|_h^2\} \end{aligned} \tag{2.19}$$

and (2.15) yields

$$\|y_n\|_{2h}^2 + \|z_h\|_h^2 \leqq 8c_1^2|f|^2. \tag{2.20}$$

Finally, using (2.18) and (2.20),

$$|z_h|^2 \leqq 8h^2c_1^2|f|^2. \tag{2.21}$$

We have thus proved the following analogue of (1.7).

PROPOSITION 2.1. *The incremental component* z_h *of* u_h *is small in the* L^2*-norm*:

$$|z_h| \leqq 4hc_1|f|. \tag{2.22}$$

We conclude this section by proving Lemmas 2.1–2.3.

Proof of Lemma 2.1. We must show that

$$\int_\Omega \nabla_h y_h \nabla_h z_h \, dx \leqq \frac{\sqrt{3}}{2}\left(\int_\Omega |\nabla_h y_h|^2 \, dx\right)^{1/2}\left(\int_\Omega |\nabla_h z_h|^2 \, dx\right)^{1/2}, \tag{2.23}$$

where $\nabla_h = (\nabla_{1h}, \nabla_{2h})$. It suffices to show (2.23) with Ω replaced by a typical coarse grid square $\mathcal{R}$ as in Fig. 2.1; (2.22) would then follow by summation for the different $\mathcal{R}$'s and utilization of the Schwarz inequality in the integrals.

For the sake of simplicity in the notation we set $y_h(M_i) = m_i$ and $z_h(A_i) = p_i$, and recall that $z_h(M_i) = 0$ for all i (see (2.11)). The function y_h is constant on the square $\mathcal{R}$ of Fig. 2.1(a) and its value is m_1. The function z_h is constant on each of the four

subsquares $\mathcal{R}_1, \cdots, \mathcal{R}_4$ numbered as shown in Fig. 2.1(b); on $\mathcal{R}_1$, $z_h = 0$, on $\mathcal{R}_2$, $z_h = p_1$, on $\mathcal{R}_3$, $z_h = p_2$, and on $\mathcal{R}_4$, $z_h = p_3$.

Now the computation is straightforward:

—On $\mathcal{R}_1$, $\nabla_h y_h = 0$, $\nabla_h z_h = \frac{1}{h}(p_1, p_2)$,

—On $\mathcal{R}_2$, $\nabla_h y_h = \frac{1}{h}(m_2 - m_1, 0)$, $\nabla_h z_h = \frac{1}{h}(-p_1, p_3 - p_1)$,

—On $\mathcal{R}_3$, $\nabla_h y_h = \frac{1}{h}(0, m_3 - m_1)$, $\nabla_h z_h = \frac{1}{h}(p_3 - p_2, -p_2)$,

—On $\mathcal{R}_4$, $\nabla_h y_h = \frac{1}{h}(m_2 - m_1, m_3 - m_1)$, $\nabla_h z_h = \frac{1}{h}(p_4 - p_3, p_5 - p_3)$.

Then

$$\int_{\mathcal{R}} |\nabla_h y_h|^2 \, dx = 2\{(m_2 - m_1)^2 + (m_3 - m_1)^2\}, \tag{2.24}$$

$$\int_{\mathcal{R}} |\nabla_h z_h|^2 \, dx = 2(p_1^2 + p_2^2) + (p_3 - p_1)^2 + (p_3 - p_2)^2 + (p_4 - p_3)^2 + (p_5 - p_3)^2, \tag{2.25}$$

$$\begin{aligned}
&\int_{\mathcal{R}} \nabla_h y_h \nabla_h z_h \, dx \\
&= -p_1(m_2 - m_1) - p_2(m_3 - m_1) + (m_2 - m_1)(p_4 - p_3) + (m_3 - m_1)(p_5 - p_3) \\
&= (m_2 - m_1)(p_4 - p_3 - p_1) + (m_3 - m_1)(p_5 - p_3 - p_2) \\
&\leqq \{2(m_2 - m_1)^2 + 2(m_3 - m_1)^2\}^{1/2} \cdot \left\{\frac{1}{2}(p_4 - p_3 - p_1)^2 + \frac{1}{2}(p_5 - p_3 - p_2)^2\right\}^{1/2} \\
&\leqq \left(\int_{\mathcal{R}} |\nabla_h y_h|^2 \, dx\right)^{1/2} \cdot \left\{\frac{3}{4}(p_4 - p_3)^2 + \frac{3}{4}(p_5 - p_3)^2 + \frac{3}{2}p_1^2 + \frac{3}{2}p_2^2\right\}^{1/2} \\
&\leqq \frac{\sqrt{3}}{2}\left(\int_{\mathcal{R}} |\nabla_h y_h|^2 \, dx\right)^{1/2} \left(\int_{\mathcal{R}} |\nabla_h z_h|^2 \, dx\right)^{1/2}.
\end{aligned}$$

Proof of Lemma 2.2. We use the same notation as in Lemma 2.1 and observe that $\nabla_{2h} y_h$ is constant in $\mathcal{R}$ and equal to $(1/2h)(m_2 - m_1, m_3 - m_1)$. Hence

$$\int_{\mathcal{R}} |\nabla_{2h} y_h|^2 \, dx = (m_2 - m_1)^2 + (m_3 - m_1)^2,$$

and it suffices to compare this to (2.24).

Proof of Lemma 2.3. Using the notation of Lemma 2.1, it suffices to prove that

$$\int_{\mathcal{R}} (z_h)^2 \, dx \leqq h^2 \int_{\mathcal{R}} |\nabla_h z_h|^2 \, dx. \tag{2.26}$$

The right-hand side of this inequality is given by (2.25) and the left-hand side is equal to

$$\begin{aligned}
h^2(p_1^2 + p_2^2 + p_3^2) &\leqq h^2\{2p_1^2 + 2p_2^2 + (p_3 - p_1)^2 + (p_3 - p_2)^2\} \\
&\leqq h^2 \int_{\mathcal{R}} |\nabla_h z_h|^2 \, dx.
\end{aligned}$$

Remark 2.2. Although here we emphasize the case $\Omega = (0, 1)^2$, let us observe that the framework and results extend to the case where Ω is any bounded domain of $\mathbb{R}^2$. In this case the mesh h can be any vector (h_1, h_2) of $\mathbb{R}^2$, $h_i > 0$, and we consider the mesh $\mathcal{R}_h$ consisting of the points $jh = (j_1 h_1, j_2 h_2)$, $j_i \in \mathbb{Z}$. We denote by $\sigma_h(M)$ the rectangle centered at M of edges h_1, h_2, and

$$\sigma_h^1(M) = \sigma_h(M) \cup \sigma_h(M + h_1 e_1) \cup \sigma_h(M + h_2 e_2).$$

Then

$$\mathring{\Omega}_h = \{M \in \mathcal{R}_h, \sigma_h^1(M) \subset \Omega\},$$
$$\Omega_h = \{M \in \mathcal{R}, \sigma_h(M) \subset \Omega\},$$

and for $M \in \Omega_h$, w_{hM} is the characteristic function of $\sigma_h(M)$. We define V_h as the space spanned by the w_{hM}, $M \in \mathring{\Omega}_h$, and $V_{2h} \subset V_h$ is defined in the same way for $2h = (2h_1, 2h_2)$. The finite difference operators ∇_{ih} are defined by

$$\nabla_{ih}\varphi(x) = \frac{1}{h_i}(\varphi(x + h_i e_i) - \varphi(x)).$$

We still have (2.5) with W_h defined exactly as above. All the results extend without any modification. Lemmas 2.1 and 2.2 are still valid, and in Lemma 2.3 we replace (2.18) by

$$\begin{aligned} |z_h| &\leqq \max(h_1, h_2)\|z_h\|_h \\ &\leqq (h_1^2 + h_2^2)^{1/2}\|z_h\|_h, \end{aligned} \tag{2.27}$$

and (2.22) is modified accordingly:

$$|z_h| \leqq 2 \max(h_1, h_2) c_1 |f|. \tag{2.28}$$

3. A class of nonlinear evolution equations. Let H be a Hilbert space endowed with the scalar product $(\cdot, \cdot)$ and the norm $|\cdot|$. We consider an evolution equation of the form

$$\frac{du}{dt} + Au + R(u) = 0 \tag{3.1}$$

with

$$R(u) = B(u) + C(u) + f. \tag{3.2}$$

The unknown function u is a map from $\mathbb{R}_+$ (or some interval of $\mathbb{R}$) into H. The operator A is linear self-adjoint unbounded in H with domain $D(A)$. We assume that A is positive closed and that A^{-1} is compact. The powers A^s of A for $S \in \mathbb{R}$ are defined and map $D(A^s)$ into H, and $D(A^s)$ is a Hilbert space for the norm $|A^s \cdot|$. We set $V = D(A^{1/2})$ and we denote by $\|\cdot\| = |A^{1/2} \cdot|$ the norm on V; $\nu > 0$ is given.

The nonlinear term $R(u)$ satisfies (3.2), where $B(u) = B(u, u)$; $B(\cdot, \cdot)$ is a bilinear continuous operator from $V \times V$ into V'; C is a linear operator from V into H and $f \in H$. We denote by b the trilinear continuous form on V given by

$$b(u, v, w) = \langle B(u, v), w \rangle \quad \forall u, v, w \in V,$$

and we assume that

$$b(u, v, v) = 0 \quad \forall u, v \in V, \tag{3.3}$$

$$|b(u, v, w)| \leqq c_2 |u|^{1/2}\|u\|^{1/2}\|v\||w|^{1/2}\|w\|^{1/2} \quad \forall u, v, w \in V, \tag{3.4}$$

$$|Cu| \leqq c_3\|u\| \quad \forall u \in V, \tag{3.5}$$

where c_2, c_3 like the quantities c_i appearing subsequently are positive constants. In addition, we require that B maps $V \times D(A)$ into H and

$$|B(u, v)| \leqq c_4 |u|^{1/2} \|u\|^{1/2} \|v\|^{1/2} |Av|^{1/2} \quad \forall u, v \in D(A), \tag{3.6}$$

$$|B(u, v)| \leqq c_5 |u|^{1/2} |Au|^{1/2} \|v\| \quad \forall u, v \in D(A). \tag{3.7}$$

Finally, we require $\nu A + C$ to be positive, i.e., there exists $\alpha > 0$ such that

$$((\nu A + C)u, u) \geqq \alpha \|u\|^2 \quad \forall u \in V. \tag{3.8}$$

Under the hypotheses above, we infer from classical results that the initial value problem consisting of (3.1) and

$$u(0) = u_0 \tag{3.9}$$

has a unique solution $u = u(t)$ defined for all $t > 0$ and such that

$$u \in \mathscr{C}(\mathbb{R}_+; H) \cap L^2(0, T; V) \quad \forall T > 0. \tag{3.10}$$

Moreover, if $u_0 \in V$, then

$$u \in \mathscr{C}(\mathbb{R}_+; V) \cap L^2(0, T; D(A)) \quad \forall T > 0. \tag{3.11}$$

Remark 3.1. As usual, we can rewrite (3.1), (3.2) in a weak (variational) form that will be appropriate for the treatment below, namely,

$$\frac{d}{dt}(u, v) + a(u, v) + b(u, u, v) + (Cu, v) = (f, v) \quad \forall v \in V, \tag{3.12}$$

where

$$a(u, v) = (Au, v). \tag{3.13}$$

We may assume for simplicity that

$$a(u, v) = \nu((u, v)), \qquad \nu > 0, \tag{3.14}$$

but this is not essential.

The abstract equation considered here includes, in particular, several dissipative evolution equations and the two-dimensional Navier–Stokes equations; however, we do not want to consider such a complicated equation here, and we will now give some simpler equations satisfying these hypotheses.

Example 1. Let Ω be an open-bounded set in $\mathbb{R}^2$ with boundary $\partial\Omega$. We consider the evolution problem (of Burgers type):

$$\frac{\partial u}{\partial t} - \Delta u + a_1 \frac{\partial u}{\partial x_1} + a_2 \frac{\partial u}{\partial x_2} + u \frac{\partial u}{\partial x_1} = f, \quad x \in \Omega, \quad t > 0, \tag{3.15}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{3.16}$$

$$u(x, 0) = u_0(x), \qquad x \in \Omega. \tag{3.17}$$

Here a_1, a_2 are given in $L^\infty(\Omega)$ and $f \in L^2(\Omega)$. We take $H = L^2(\Omega)$, $V = H_0^1(\Omega)$, $A = -\Delta$ with Dirichlet boundary conditions:

$$B(u, v) = u \frac{\partial v}{\partial x_1}, \qquad Cu = a_1 \frac{\partial u}{\partial x_1} + a_2 \frac{\partial u}{\partial x_2}.$$

It is easy to check that all the hypotheses are satisfied.

Example 2. This is an example of an integral Burgers type equation. Everything else being unchanged, we now consider

$$B(u, v) = \left(\int_\Omega u(\xi)\, d\xi \right) \frac{\partial v}{\partial x_1}.$$

All the hypotheses above are satisfied.

Example 3. This last example is a system close to the Navier-Stokes equations but without the pressure term and the incompressibility condition; and the nonlinear (inertial) term is modified as in [T1]. Let Ω be an open-bounded set of $\mathbb{R}^2$ with boundary $\partial\Omega$. The function $u = (u_1, u_2)$ maps $\Omega \times (0, T)$ into $\mathbb{R}^2$ and satisfies

$$\frac{\partial u}{\partial t} - \nu \Delta u + (u \cdot \nabla) u + \frac{1}{2} (\operatorname{div} u) u = f, \quad x \in \Omega, \quad t > 0, \tag{3.18}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{3.19}$$

$$u(x, 0) = u_0(x). \tag{3.20}$$

We take $H = L^2(\Omega)^2$, $V = H_0^1(\Omega)^2$, $D(A) = \{H_0^1(\Omega) \cap H^2(\Omega)\}^2$, $A = -\nu\Delta$, $C = 0$, and

$$B(u, v) = (u \cdot \nabla) v + \tfrac{1}{2} (\operatorname{div} u) v.$$

All the hypotheses are satisfied.

4. Spatial discretization. The IMG algorithm. In this section we present the IMG algorithm, which combines the ideas and concepts on inertial manifolds and those on multigrid methods. We start by describing the spatial discretization of (3.1) (or (3.12)) in a way suitable for our purpose. We then give an estimate on the incremental component of the solution, which is a partial justification for the IMG algorithm. Finally, we describe the IMG algorithm itself.

4.1. Spatial discretization. For the spatial discretization of (3.1), (3.12), we are traditionally given a family of finite-dimensional spaces $V_h \subset H$ endowed with a Hilbert scalar product and norm $((\cdot, \cdot))h$, $\|\cdot\|_h$. The parameter h is a discretization parameter. For Galerkin methods, and particularly for finite-element methods, $V_h \subset V$ and it is required that $\cup_h V_h$ is dense in V. For finite differences, as shown in § 2, the space V_h is not a subspace of V and the hypotheses are slightly more involved; they will not be recalled here (see, however, § 5).

For the spatial discretization of (3.12), we then consider a function u_h from $\mathbb{R}_+$ into V_h, which satisfies

$$\frac{d}{dt}(u_h, v_h) + a_h(u_h, v_h) + b_h(u_h, u_h, v_h) + (C_h u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h, \tag{4.1}$$

where a_h, v_h, and C_h are appropriate approximations of a, v, and C. Of course, the second step in the numerical approximation of (3.1), (3.12) is the time discretization of (4.1), but we emphasize the spatial discretization here.

4.2. Incremental unknowns. As in § 2, we now consider two values of the discretization parameter, h and $2h$, and the corresponding approximating spaces V_h and V_{2h}. It is assumed that $V_{2h} \subset V_h$ and, more precisely, we write as in (3.5)

$$V_h = V_{2h} \oplus W_h. \tag{4.2}$$

Any $u_h \in V_h$ is decomposed into

$$u_h = y_h + z_h, \tag{4.3}$$

where y_h is called the coarse grid component of u_h and z_h is called the incremental component of u_h. We assume the following properties, which have been proved in § 2 when $V = H_0^1(\Omega)$ and $\Omega = (0, 1)^2$:

$$|((y_h, z_h))_h| \leqq (1-\delta)\|y_h\|_h \|z_h\|_h \quad \forall y_h \in V_{2h}, \quad \forall z_h \in W_h, \tag{4.4}$$

where $0 < \delta < 1$ is independent of h, and

$$S_1(h)|z_h| \leqq \|z_h\|_h \quad \forall z_h \in W_h, \tag{4.5}$$

where $S_1(h) \to \infty$ as $h \to 0$.

We now use the decomposition (4.3) of u_h and it is straightforward that (4.1) is equivalent to the following system for y_h, z_h:

$$\left(\frac{d}{dt}(y_h + z_h), \tilde{y}_h\right) + a_h(y_h + z_h, \tilde{y}_h) + b_h(y_h + z_h, y_h + z_h, \tilde{y}_h) + (C_h(y_h + z_h), \tilde{y}_h) = (f, \tilde{y}_h) \quad \forall \tilde{y}_h \in V_{2h}, \tag{4.6}$$

$$\left(\frac{d}{dt}(y_h + z_h), \tilde{z}_h\right) + a_h(y_h + z_h, \tilde{z}_h) + b_h(y_h + z_h, y_h + z_h, \tilde{z}_h) + (C_h(y_h + z_h), \tilde{z}_h) = (f, \tilde{z}_h) \quad \forall \tilde{z}_h \in W_h. \tag{4.7}$$

This is just a rewriting of the natural spatial discretization scheme (4.1) corresponding to V_h. However, as a partial justification for the IMG algorithm described below, we want to show now that the incremental component z_h is small for h small.

4.3. Estimate on the incremental component. To derive a priori estimates we must assume that a_h, b_h, C_h satisfy some hypotheses similar to those on a, b, C (see (3.3)-(3.8)). We assume here that

$$b_h(u_h, v_h, v_h) = 0 \quad \forall u_h, v_h \in V_h, \tag{4.8}$$

$$|b_h(u_h, v_h, w_h)| \leqq c_6 |u_h|^{1/2} \|u_h\|_h^{1/2} \|v_h\| \, |w_h|^{1/2} \|w_h\|_h^{1/2} \quad \forall u_h, v_h, w_h \in V_h, \tag{4.9}$$

$$|a_h(u_h, v_h)| \leqq c_7 \|u_h\|_h \|v_h\|_h, \qquad |C_h u_h| \leqq c_8 \|u_h\|_h \quad \forall u_h, v_h \in V_h, \tag{4.10}$$

$$a_h(u_h, u_h) + (C_h u_h, u_h) \geqq \alpha_1 \|u_h\|_h^2 \quad \forall u_h \in V_h, \tag{4.11}$$

$$|u_h|_h \leqq c_9 \|u_h\|_h \quad \forall u_h \in v_h, \tag{4.12}$$

where c_6-c_9 and α_1 are independent of h ($\alpha_1 > 0$).

As usual, a priori estimates on u_h are obtained by replacing v_h by $u_h (= u_h(t))$ in (4.1). It is equivalent to replacing $\tilde{y}_h$ by $y_h (= y_h(t))$ in (4.6) and $\tilde{z}_h$ by $z_h (= z_h(t))$ in (4.7) and adding the equations that we obtain. Thanks to (4.8) we find

$$\frac{1}{2}\frac{d}{dt}|u_h|^2 + a_h(u_h, u_h) + (C_h u_h, u_h) = (f, u_h).$$

Using (4.11) and (4.12), we can then write

$$\frac{1}{2}\frac{d}{dt}|u_h|^2+\alpha_1\|u_h\|_h^2\leqq|f||u_h|$$

$$\leqq c_9|f|\,\|u_h\|_h$$

$$\leqq\frac{\alpha_1}{2}\|u_h\|_h^2+\frac{c_9^2}{2\alpha_1}|f|^2,$$

$$\frac{d}{dt}|u_h|^2+\alpha_1\|u_h\|_h^2\leqq\frac{c_9^2}{\alpha_1}|f|^2, \tag{4.13}$$

$$\frac{d}{dt}|u_h|^2+\frac{\alpha_1}{c_8^2}|u_h|^2\leqq\frac{c_9^2}{\alpha_1}|f|^2. \tag{4.14}$$

The last inequality yields a uniform estimate of u_h in $\mathscr{C}(\mathbb{R}_+;H)$:

$$|u_h(t)|^2\leqq|u_h(0)|^2\exp\left(-\frac{\alpha_1}{c_9^2}t\right)+\frac{c_8^2c_9^2}{\alpha_1^2}|f|^2\quad\forall t>0. \tag{4.15}$$

Returning to (4.13) we find, for any $T>0$

$$\frac{1}{T}\int_0^T\|u_h\|^2\,dt\leqq\frac{1}{T\alpha_1}|u_h(0)|^2+\frac{c_9^2}{\alpha_1^2}|f|^2\leqq K_1. \tag{4.16}$$

Separate estimates can be obtained for y_h and z_h by using (4.4). Indeed,

$$\|u_h\|^2=\|y_h+z_h\|_h^2$$

$$=\|y_h\|_h^2+\|z_h\|_h^2+2((y_h,z_h))_h \tag{4.17}$$

$$\geqq 2\delta(\|y_h\|_h^2+\|z_h\|_h^2),$$

and hence

$$\frac{1}{T}\int_0^T(\|y_h\|_h^2+\|z_h\|_h^2)\,dt\leqq\frac{K_1}{2\delta}. \tag{4.18}$$

Then, using (4.5), we obtain that z_h is small in the following sense (at least):

$$\frac{1}{T}\int_0^T|z_h|^2\,dt\leqq\frac{K_1}{2\delta}(S_1(h))^{-2}. \tag{4.19}$$

We can derive further a priori estimates on z_h, but they involve more complicated computations and will not be given here.

4.4. The IMG algorithm. The algorithm that we now present stems from the the theory of dynamical systems, the idea being to approximate the universal attractor describing the long-term behavior of the solutions of (3.1), (3.2). A partial justification of this algorithm lies in the fact that some of the terms involving z_h in (4.6), (4.7) are small and thus can be neglected. We refer the reader to [FMT] and [T5], [T6] for further justification.

The algorithm that we consider is the following: $u_h=y_h+z_h$ is an approximation of u different from that above and, in particular, y_h, z_h are no longer the same as in (4.6), (4.7).

We look for $u_h = y_h + z_h$, satisfying

$$\left(\frac{dy_h}{dt}, \tilde{y}_h\right) + a_h(y_h + z_h, \tilde{y}_h) + b_h(y_h, y_h, \tilde{y}_h) + b_h(y_h, z_h, \tilde{y}_h) + b_h(z_h, y_h, \tilde{y}_h) + (C_h(y_h + z_h), \tilde{y}_h) = (f, \tilde{y}_h) \quad \forall \tilde{y}_h \in V_{2h}, \tag{4.20}$$

$$a_h(y_h + z_h, \tilde{z}_h) + b_h(y_h, y_h, \tilde{z}_h) + (C_h(y_h + z_h), \tilde{z}_h) = (f, \tilde{z}_h) \quad \forall \tilde{z}_h \in W_h, \tag{4.21}$$

$$(y_h(0), \tilde{y}_h) = (u_0, \tilde{y}_h) \quad \forall \tilde{y}_h \in V_{2h}. \tag{4.22}$$

Theoretical questions including the existence of y_h, z_h will be studied in § 5. For the moment we conclude this section with some general comments.

We rewrite (4.21) in the form

$$a_h(z_h, \tilde{z}_h) + (C_h z_h, \tilde{z}_h) = -a_h(y_h, \tilde{z}_h) - b_h(y_h, y_h, \tilde{z}_h) - (C_h y_h, \tilde{z}_h) + (f, \tilde{z}_h) \quad \forall \tilde{z}_h \in W_h. \tag{4.23}$$

It then follows from the Lax-Milgram Theorem and hypothesis (4.11) that, at each instant of time, $z_h = z_h(t)$ is uniquely determined as a function of $y_h = y_h(t)$ and of the other data:

$$z_h(t) = \Phi_h(y_h(t)). \tag{4.24}$$

By inserting this expression of z_h in (4.20), we find the following equation for y_h:

$$\left(\frac{dy_h}{dt}, \tilde{y}_h\right) + a_h(y_h + \Phi_h(y_h), \tilde{y}_h) + b_h(y_h, y_h, \tilde{z}_h) + b_h(y_h, \Phi_h(y_h), \tilde{z}_h) + b_h(\Phi_h(y_h), y_h, \tilde{y}_h) + (C_h(y_h + \Phi_h(y_h)), \tilde{y}_h) = (f, \tilde{y}_h) \quad \forall \tilde{y}_h \in V_{2h}. \tag{4.25}$$

We observe that, setting $\Phi_h = 0$ in (4.25), (4.22), we recover exactly the approximation based on V_{2h}. Hence (4.25) gives a V_{2h} approximation of y_h, perturbed by some "small terms" $z_h = \Phi_h(y_h)$. The solution u_h of the IMG algorithm lies on the manifold $\mathcal{M}_h$ of V_h of equation

$$z_h = \Phi_h(y_h). \tag{4.26}$$

The dimension of this manifold is that of V_{2h}. Note that, with a usual V_{2h} discretization ($z_h = 0$), the solution $u_h = y_h$ would lie in V_{2h}; and with a V_h-discretization as in (4.1) (or (4.6), (4.7)), the solution u_h could be a priori anywhere in V_h. Here the solution lies on the manifold $\mathcal{M}_h$ of V_h whose dimension is that of V_{2h}. As in [T5] for the case of spectral methods, it is expected that the IMG algorithm provides a V_h-accuracy, with a V_{2h}-complexity of computation.

Remark 4.1. Although we restricted ourselves to two discretization meshes h and $2h$, we could as well consider two discretization meshes h and dh, $d \in \mathbb{N}$ fixed ($= 2, 3, 4, \cdots$). All the developments above would be still valid without any modification; we only need to extend (4.4), (4.15) (i.e., Lemmas 2.1 and 2.3) to this case.

5. Convergence of the algorithm. Our aim is now to prove the convergence of the algorithm presented in § 4. We do not make any attempt at generality and, strictly speaking, the convergence result hereafter applies to the three examples described in § 3. However, the methodology is general, and with appropriate hypotheses—including in particular the so-called consistency hypotheses that specify how a_h, b_h, C_h, v_h approximate a, b, C, V—the result applies to the general equation (3.1), (3.12) and to even more general situations.

THEOREM 5.1. *The hypotheses are those above, u_0 is given in H, and $u = u(t)$ is the solution of* (3.1), (3.9). *For every fixed h, the solution $u_h = y_h + z_h$ of* (4.20)-(4.22) *exists and is uniquely defined for all $t > 0$. When $h \to 0$, u_h converges to u in the following sense:*

(5.1) *$y_h \to u$ in $L^p(0, T; H)$ strongly for all $T > 0$ and all p, $1 \leqq p < \infty$, $\nabla_h u_h \to \nabla u$ in $L^2(\Omega \times (0, T))$ strongly, for all $T > 0$.*

(5.2) *$z_h \to 0$ in $L^p(0, T; H)$ strongly for all $T > 0$ and all p, $1 \leqq p < \infty$, $\nabla_h z_h \to 0$ in $L^2(\Omega \times (0, T))$ strongly, for all $T > 0$.*

(5.3) *$y_h \to u$ and $z_h \to 0$ in $L^\infty(\mathbb{R}_+; H)$ weak-star.*

The proof of Theorem 5.1 is given below. It comprises several steps, the first ones being devoted to the derivation of a priori estimates.

5.1. A priori estimates I. We start by deriving a priori estimates for y_h and z_h. They are similar to those derived in (4.18), (4.19) for the usual discretization procedure.

The system of equations (4.20)-(4.22) is equivalent to (4.25) and (4.22), the expression of Φ_h being given by (4.24). Since Φ_h is a simple (quadratic) function, the existence of y_h (and thus z_h) on some interval of time $[0, T_h)$ follows readily from classical theorems on differential equations. The fact that y_h and z_h are defined for all $t > 0$ (i.e., $T_h = +\infty$) will follow from the a priori estimates below.

We replace $\tilde{y}_h$ by y_h $(= y_h(t))$ in (4.20) and $\tilde{z}_h$ by z_h $(= z_h(t))$ in (4.21) and we add the equations that we obtain. We find

$$\frac{1}{2}\frac{d}{dt}|y_h|^2 + a_h(y_h + z_h, y_h + z_h) + (C_h(y_h + z_h), y_h + z_h) = (f, y_h + z_h). \tag{5.4}$$

We have used (4.8) and its consequence

$$b_h(\varphi_h, \psi_h, \theta_h) = -b_h(\varphi_h, \theta_h, \psi_h) \quad \forall \varphi_h, \psi_h, \theta_h \in V_h. \tag{5.5}$$

Thanks to (4.11) we then have

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}|y_h|^2 + \alpha_1\|y_h + z_h\|_h^2 &= (f, y_h + z_h) \\
&\leqq |f||y_h + z_h| \quad \text{(by the Schwarz inequality and (4.12))} \\
&\leqq c_9|f|\,\|y_h + z_h\|_h \\
&\leqq \frac{\alpha_1}{2}\|y_h + z_h\|_h^2 + \frac{c_9^2}{2\alpha_1}|f|^2,
\end{aligned}$$

$$\frac{d}{dt}|y_h|^2 + \alpha_1\|y_h + z_h\|_h^2 \leqq \frac{c_9^2}{\alpha_1}|f|^2. \tag{5.6}$$

Using (4.4) as in (4.17) and again using (4.11), we obtain

$$\frac{d}{dt}|y_h|^2 + 2\alpha_1\delta(\|y_h\|_h^2 + \|z_h\|_h^2) \leqq \frac{c_9^2}{\alpha_1}|f|^2, \tag{5.7}$$

$$\frac{d}{dt}|y_h|^2 + \frac{2\alpha_1\delta}{c_9^2}|y_h|^2 \leqq \frac{c_9^2}{\alpha_1}|f|^2. \tag{5.8}$$

We infer from (5.8) and from Gronwall's Lemma the following estimate on y_h:

$$|y_h(t)|^2 \leqq |y_h(0)|^2 \exp\left(-\frac{2\alpha_1\delta}{c_9^2}t\right) + \frac{c_9^4}{2\alpha_1^2\delta}|f|^2. \tag{5.9}$$

This shows that y_h remains bounded, and y_h is defined for all $t>0$ (i.e., $T_h=+\infty$). Furthermore, since by (4.22)

$$|y_h(0)|\leqq|u_0|, \tag{5.10}$$

we have a time-uniform estimate for y_h that is independent of h:

(5.11) y_h remains in a bounded set of $L^\infty(\mathbb{R}_+;H)$ as $h\to 0$.

As we have already mentioned, (4.23) and (4.24) imply that z_h too is defined for all $t>0$. By integration of (5.7) between zero and T ($T>0$ fixed), we then deduce that

$$\frac{1}{T}\int_0^T(\|y_h\|_h^2+\|z_h\|_h^2)\,dt\leqq\frac{1}{2\alpha_1\delta T}|y_h(0)|^2+\frac{c_9^2}{2\alpha_1\delta T}|f|^2. \tag{5.12}$$

This implies that

(5.13) For every $T>0$ fixed, the norms of y_h and z_h in $L^2(0,T;V_h)$ remain bounded as $h\to 0$.

By use of (4.5) this yields

(5.14) For every $T>0$ fixed, the norm of $\{S_1(h)\}z_h$ in $L^2(0,T;H)$ remains bounded as $h\to 0$.

Remark 5.1. We note that due to the term $1/T$ in front of the integral in the left-hand side of (5.12), this inequality yields slightly more than (5.13). More generally, we can integrate (5.7) between t and $t+T$ (t, $T>0$ fixed); using (5.9) and (5.10) we then find

$$\begin{aligned}\frac{1}{T}\int_0^{t+T}(\|y_h\|_h^2+\|z_h\|_h^2)\,ds&\leqq\frac{1}{2\alpha_1\delta T}|y_h(t)|^2+\frac{c_9^2}{2\alpha_1^2\delta}|f|^2\\&\leqq\frac{1}{2\alpha_1\delta T}|y_h(0)|^2\exp\left(-\frac{2\alpha_1\delta}{c_9^2}t\right)+\left(\frac{c_9^4}{4\alpha_1^3\delta^2T}+\frac{c_9^2}{2\alpha_1^2\delta}\right)|f|^2,\end{aligned} \tag{5.15}$$

$$\frac{1}{T}\int_t^{t+T}(\|y_h\|_h^2+\|z_h\|_h^2)\,ds\leqq\frac{1}{2\alpha_1\delta T}|u_0|^2+\left(\frac{c_9^4}{4\alpha_1^3\delta^2T}+\frac{c_9^2}{2\alpha_1^2\delta}\right)|f|^2.$$

This estimate independent of h is valid for all $t>0$, $T>0$. Of course (5.15) also implies an analogue of (5.14).

Remark 5.2. We recall that in the case where V_h is the discrete analogue of $H_0^1(\Omega)$ as in § 2, i.e., for the five-point discretization of the Laplace operator, then

$$\|y_h\|_h^2=\int_\Omega|\nabla_h y_h|^2\,dx,$$

$$\int_0^T\|y_h(t)\|_h^2\,dt=\int_0^T\int_\Omega|\nabla_h y_h(x,t)|^2\,dx\,dt.$$

Therefore (5.13) is an estimate in $L^2(\Omega\times(0,T))$, independent of h, for the discrete analogue of the gradient of y_h and z_h.

5.2. A priori estimates II. The following a priori estimates give some improvements that are not essential but are useful. We first restrict ourselves to the framework of § 2 (i.e., to Examples 1–3 of § 3). Since V_h is finite-dimensional, the norms of V_h are equivalent; thus (2.14) (or (4.12)) can be supplemented by another inequality:

$$\|u_h\|_h\leqq S_2(h)|u_h|\quad\forall u_h\in V_h, \tag{5.16}$$

where $S_2(h) \to \infty$ as $h \to 0$. In the situation of § 2 the computation of $S_2(h)$ is easy and we find

$$S_2(h) = \frac{2\sqrt{2}}{h}. \tag{5.17}$$

Now we set $\tilde{z}_h = z_h$ $(= z_h(t))$ in (4.23) and (4.8)-(4.11):

$$a_h(z_h, z_h) + (C_h z_h, z_h) = -a_h(y_h, z_h) - b_h(y_h, y_h, z_h) - (C_h y_h, z_h) + (f, z_h),$$

$$\alpha_1 \|z_h\|_h^2 \leqq c_7 \|y_h\|_h \|z_h\|_h + c_8 \|y_h\|_h |z_h| + c_6 |y_h| \|y_h\|_h \|z_h\|_h + |f||z_h|$$

$$\leqq (c_7 + c_8 S_1(h)^{-1} + c_6 |y_h|) \|y_h\|_h \|z_h\|_h + S_1(h)^{-1} |f| \|z_h\|_h \quad \text{(because of (4.5))}.$$

Again using (4.5) and (5.16), we find

$$|z_h| \leqq (\alpha_1 S_1(h))^{-1} (c_7 + c_8 S_1(h)^{-1} + c_6 |y_h|) \|y_h\|_h + \frac{1}{\alpha_1} (S_1(h))^{-3} |f|$$

$$\leqq \frac{1}{\alpha_1} \frac{S_2(h)}{S_1(h)} (c_7 + c_8 S_1(h)^{-1} + c_6 |y_h|) |y_h| + \frac{1}{\alpha_1} (S_1(h))^{-2} |f|.$$

Owing to the expression of $S_1(h)$ in (2.18) $(S_1(h) = h^{-1})$ and to the expression of $S_2(h)$ in (5.17), we have

$$\frac{S_2(h)}{S_1(h)} \leqq 2\sqrt{2}. \tag{5.18}$$

Finally, thanks to (5.11), we conclude that

(5.19) z_h remains in a bounded set of $L^\infty(\mathbb{R}_+; H)$ as $H \to 0$.

5.3. Passage to the limit. The passage to the limit $h \to 0$ relies on fairly standard methods. We will only sketch this step of the proof; the reader is referred to [T2] for the details in related situations.

Thanks to (5.11), (5.13), Remark 5.2, (5.14), and (5.19), there exists a subsequence (still denoted h) and there exists u:

$$u \in L^\infty(\mathbb{R}_+; H) \cap L^2(0, T; V) \quad \forall T > 0, \tag{5.20}$$

such that for $h \to 0$

$$\begin{aligned} &y_h \to u \text{ in } L^\infty(\mathbb{R}_+; H) \text{ weak-star,} \\ &\nabla_h y_h \to \nabla u \text{ in } L^2(\Omega \times (0, T)) \text{ weakly} \quad \forall T > 0, \end{aligned} \tag{5.21}$$

$$\begin{aligned} &z_h \to 0 \text{ in } L^\infty(\mathbb{R}_+; H) \text{ weak-star,} \\ &\nabla_h z_h \to 0 \text{ in } L^2(\Omega \times (0, T)) \text{ weakly} \quad \forall T > 0. \end{aligned} \tag{5.22}$$

We infer from (5.14) and (5.19) that

$$z_h \to 0 \text{ in } L^p(0, T; H) \text{ strongly} \quad \forall 1 \leqq p < \infty, \quad \forall T > 0. \tag{5.23}$$

Also, by using a compactness argument recalled below, we can improve (5.21) and show that

$$y_h \to u \text{ in } L^p(0, T; H) \text{ strongly} \quad \forall 1 \leqq p < \infty, \quad \forall T > 0 \quad \text{as } h \to 0. \tag{5.24}$$

Using the convergences (5.21)-(5.24) we can then pass to the limit in (4.20)-(4.22), following classical methods. At the limit we find that u is solution of (3.12) (or (3.1)) and (3.9). Since the solution to this problem is unique, we see by a contradiction argument that the convergences (5.21)-(5.24) hold for the whole sequence h.

Finally, for the strong convergence of the derivatives we consider the restriction operators r_h (see [C], [T2]) that map V into V_h and such that

$$\begin{aligned} &r_h u \to u && \text{in } L^2(0, T; H) \text{ strongly,} \\ &\nabla_h r_h u \to \nabla u && \text{in } L^2(0, T; H) \text{ strongly.} \end{aligned} \tag{5.25}$$

We then consider the following expression, where $u_h = y_h + z_h$:

$$X_h = \frac{1}{2}|y_h(T) - u(T)|^2 + \int_0^T \cdot \{a_h(u_h - r_h u, u_h - r_h u) + (C_h(u_h - r_h u, (C_h(u_h - r_h u), u_h - r_h u)u_h - r_h u)\}\, dt.$$

We have $X_h = X_h^1 + X_h^2 + X_h^3$:

$$X_h^1 = \frac{1}{2}|y_h(T)|^2 + \int_0^T \{a_h(u_h, u_h) + (C_h u_h, u_h)\}\, dt$$

$$= \frac{1}{2}|y_h(0)|^2 + \int_0^T (f, u_h)\, dt \quad \text{(by (5.4))};$$

it is clear that, for $h \to 0$,

$$X_h^1 \to X^1 = \frac{1}{2}|u_0|^2 + \int_0^T (f, u)\, dt;$$

$$X_h^2 = -(y_h(T), u(T)) - 2\int_0^T \{a_h(u_h, r_h u) + (C_h u_h, r_h u)\}\, dt;$$

as $h \to 0$, X_h^2 converges to

$$X^2 = -|u(T)|^2 - 2\int_0^T \{a(u, u) + (Cu, u)\}\, dt,$$

$$X^2 = -2X^1,$$

$$X_h^3 = \frac{1}{2}|u(T)|^2 + \int_0^T \{a_h(r_h u, r_h u) + (C_h r_h u, r_h u)\}\, dt.$$

Thanks to (5.25), when $h \to 0$, X_h^3 converges to $-\frac{1}{2}X^2 = X^1$.

Finally X_h converges to zero and thanks to (4.11) we conclude that

$$\int_0^T \|u_h - r_h u\|_h^2\, dt \to 0 \quad \text{as } h \to 0.$$

We denote by $\tilde{y}_h(t)$ and $\tilde{z}_h(t)$ the components of $(r_h u)(t)$ on V_{2h} and W_h. Using again (4.4) as in (4.17), we find

$$\int_0^T \|y_h - \tilde{y}_h\|_h^2\, dt \to 0 \quad \text{as } h \to 0,$$

$$\int_0^T \|z_h - \tilde{z}_h\|_h^2\, dt \to 0 \quad \text{as } h \to 0.$$

The conclusion follows then from the easy to prove fact that the incremental component $\tilde{z}_h$ of $r_h u$ converges to zero as $h \to 0$.

The proof of Theorem 5.1 is complete after we prove (5.24).

5.4. Compactness. The strong convergence result in (5.24) is shown by using a compactness theorem. The most appropriate here is that in [T3]. To apply this compactness result, we observe (see, for instance, [T2]) that the set

$$\{\varphi_h \in H,\ \|\varphi_h\|_h \leqq 1\} \tag{5.26}$$

is relatively compact in $H = L^2(\Omega)$. Furthermore we must estimate, for $r>0$ fixed,

$$\int_0^T |y_h(t+r)-y_h(t)|^2\,dt. \tag{5.27}$$

By integration of (4.20) we find[4]

$$\begin{aligned}(y_h(t+r)-y_h(t),\tilde{y}_h)+a_h\left(\int_t^{t+r} u_h(s)\,ds,\tilde{y}_h\right)+\left(C_h\int_t^{t+r} u_h(s)\,ds,\tilde{y}_h\right)\\ +\left(\int_t^{t+r} B_h(y_h(s))+B_h(y_h(s),z_h(s))+B_h(z_h(s),y_h(s))\,ds,\tilde{y}_h\right)\\ =\left(\int_t^{t+r} f(s)\,ds,\tilde{y}_h\right).\end{aligned}$$

We then set $\tilde{y}_h = y_h(t+r)-y_h(t)$ and integrate the resulting equation with respect to t. This gives

$$\int_0^T |y_h(t+r)-y_h(t)|^2\,dt \leqq \sum_{j=1}^{6} I_j$$

with

$$\begin{aligned}I_1&=\left|\int_0^T a_h\left(\int_t^{t+r} u_h(s)\,ds, y_h(t+r)-y_h(t)\right)dt\right|\\ &\leqq r^{1/2}c_7\left(\int_0^{T+r}\|u_h(s)\|_h^2\,ds\right)^{1/2}\left(\int_0^T \|y_h(t+r)-y_h(t)\|_h\,dt\right)\\ &\leqq cr^{1/2},\end{aligned}$$

where here and below c is independent of r and h;

$$\begin{aligned}I_2&=\left|\int_0^T\left(C_h\int_t^{t+r} u_h(s)\,ds, y_h(t+r)-y_h(t)\right)dt\right|\\ &\leqq c_8 r^{1/2}\left(\int_0^{T+r}\|u_h(s)\|_h^2\,ds\right)^{1/2}\left(\int_0^T |y_h(t+r)-y_h(t)|\,dt\right)\\ &\leqq cr^{1/2},\end{aligned}$$

$$\begin{aligned}I_3&=\left|\int_0^T\left(\int_t^{t+r} B_h(y_h(s))\,ds, y_h(t+r)-y_h(t)\right)dt\right|\\ &\leqq c_6\int_0^T\left(\int_t^{t+r}|y_h(s)|\,\|y_h(s)\|_h\,ds\right)\|y_h(t+r)-y_h(t)\|_h\,dt \quad \text{(from (4.9) and (5.5))}\\ &\leqq c_6 r^{1/2}|y_h|_{L^\infty(\mathbb{R}_+;H)}\left(\int_0^T\|y_h(s)\|_h^2\,ds\right)^{1/2}\cdot\left(\int_0^T\|y_h(t+r)-y_h(t)\|_h\,dt\right)\\ &\leqq cr^{1/2}.\end{aligned}$$

[4] $B_h(\varphi_h)=B_h(\varphi_h,\varphi_h)$, where $B_h(\cdot,\cdot)$ is the bilinear mapping from $V_h\times V_h$ into V_h defined by $(B_h(\varphi_h,\psi_h),\theta_h)=b_h(\varphi_h,\psi_h,\theta_h)$, for all $\varphi_h,\psi_h,\theta_h\in V_h$.

The two other terms involving B_h and denoted I_4, I_5 are estimated similarly. Finally,

$$I_6 = \left| \int_0^T \left(\int_t^{t+r} f(s)\, ds, y_h(t+r) - y_h(t) \right) dt \right|$$
$$\leqq r^{1/2} \left(\int_0^{T+r} |f(s)|^2\, ds \right)^{1/2} \left(\int_0^T |y_h(t+r) - y_h(t)|_h\, dt \right)$$
$$\leqq c r^{1/2}.$$

In conclusion (5.27) is bounded by $cr^{1/2}$ and tends to zero as $r \to 0$, uniformly with respect to h. As is shown in [T3] this, together with (5.26), ensures (5.24).

Remark 5.3. The extension of Theorem 5.1 to more general examples than those of § 3 necessitate the following hypotheses:

(i) $S_2(h)/S_1(h)$ is bounded by a constant independent of h (see (5.16) and (5.18)).

(ii) The consistency hypotheses specifying how a_h, b_h, C_h, and V_h approximate a, b, C, and V.

(iii) The definition of a restriction operator r_h mapping V into V_h and such that (5.25) holds.

For example, in the case of the Dirichlet problem in a bounded domain Ω of $\mathbb{R}^2$ with meshes h_1, h_2 different in both directions, as in Remark 2.1, then

$$S_1(h) = 2c_1 \{\max(h_1, h_2)\}^{-1}$$

(see Remark 2.1); an easy computation shows that

$$S_2(h) = \sqrt{2} \left(\frac{1}{h_1^2} + \frac{1}{h_2^2} \right)^{1/2}.$$

Hence (i) is satisfied if h_1/h_2 remains bounded from above and below.

6. Improved stability. One of the advantages produced by the IMG algorithm is an improved stability condition when discretizations in space and time are both performed and an explicit time-discretization scheme is used. Since this improvement is already transparent in the linear case we will restrict ourselves to the linear case and replace b, C, b_h, C_h by 0. Without any loss of generality we can also set $f = 0$; finally for simplicity we take

$$\begin{aligned} &a(u, v) = \nu((u, v)), \qquad \nu > 0, \\ &a_h(u_h, v_h) = \nu((u_h, v_h))_h. \end{aligned} \tag{6.1}$$

After discretization in time by an explicit scheme the IMG algorithm (4.20)-(4.22) now leads to the construction of two sequences of elements of V_h:

$$\begin{aligned} &y_h^n \in V_{2h}, && n \geqq 0, \\ &z_h^n \in W_h, && n \geqq 0, \\ &u_h^n = y_h^n + z_h^n \in V_h, && n \geqq 0. \end{aligned}$$

These elements are defined recursively. We first define y_h^0 by setting as in (4.22):

$$(y_h^0, \tilde{y}_h) = (u_0, \tilde{y}_h) \quad \forall \tilde{y}_h \in V_{2h}. \tag{6.2}$$

When y_h^n is known, $n \geqq 0$, we define z_h^n and y_h^{n+1} as follows:

$$\frac{1}{k}(y_h^{n+1} - y_h^n, \tilde{y}_h) + a_h(y_h^n + z_h^n, \tilde{y}_h) = 0 \quad \forall \tilde{y}_h \in V_{2h}, \tag{6.3}$$

$$a_h(y_h^n + z_h^n, \tilde{z}_h) = 0 \quad \forall \tilde{z}_h \in W_h. \tag{6.4}$$

Here $T>0$ is fixed, $M\in\mathbb{N}$ is given and $k=T/M$ is the time discretization mesh. We rewrite (6.4) as

$$a_h(z_h^n,\tilde{z}_h)=-a_h(y_h^n,\tilde{z}_h)\quad\forall\tilde{z}_h\in W_h \tag{6.5}$$

and we infer easily the existence and uniqueness of $z_h^n\in W_h$ satisfying (4.3)-(4.4) from the Lax-Milgram Theorem. Once z_h^n (and y_h^n) are known, (6.3) readily determines $y_h^{n+1}\in V_{2h}$. The construction can then continue.

The stability condition arises when we try to determine a priori estimates on the sequences y_h^n, z_h^n. For the a priori estimates we replace $\tilde{y}_h$ by y_h^n in (6.3) and $\tilde{z}_h$ by z_h^n in (6.4). We have

$$2(y_h^{n+1}-y_h^n,y_h^n)=|y_h^{n+1}|^2-|y_h^n|^2-|y_h^{n+1}-y_h^n|^2.$$

Thus,

$$|y_h^{n+1}|^2-|y_h^n|^2-|y_h^{n+1}-y_h^n|^2+2k\nu((u_h^n,y_h^n))_h=0, \tag{6.6}$$

$$\nu((u_h^n,z_h^n))_h=0, \tag{6.7}$$

and by adding (6.6) to $2k$ times (6.7) we find

$$|y_h^{n+1}|^2-|y_h^n|^2+2k\nu\|u_h^n\|_h^2=|y_h^{n+1}-y_h^n|^2. \tag{6.8}$$

To estimate the right-hand side of (6.8) we replace $\tilde{y}_h$ by $k(y_n^{n+1}-y_h^n)$ in (6.3) and we obtain

$$\begin{aligned}|y_h^{n+1}-y_h^n|^2&=-k\nu((u_h^n,y_h^{n+1}-y_h^n))_h\\&\leqq k\nu\|u_h^n\|_h\|y_h^{n+1}-y_h^n\|_h.\end{aligned}$$

Now we can use *the analogue of* (5.16) *in* V_{2h}:

$$\|\varphi_h\|_h\leqq S_2(2h)|\varphi_h|\quad\forall\varphi_h\in V_{2h},$$

and this leads to

$$\begin{aligned}&|y_h^{n+1}-y_h^n|^2\leqq k\nu S_2(2h)\|u_h^n\|_h|y_h^{n+1}-y_h^n|,\\&|y_h^{n+1}-y_h^n|^2\leqq(k\nu S_2(2h))^2\|u_h^n\|_h^2.\end{aligned} \tag{6.9}$$

We compare (6.8) and (6.9) and write:

$$|y_h^{n+1}|^2-|y_h^n|^2+2k\nu(1-\tfrac{1}{2}k\nu(S_2(2h))^2)\|u_h^n\|_h^2\leqq 0\quad\forall n\geqq 0. \tag{6.10}$$

We can deduce the desired a priori estimates from (7.10) provided k, h satisfy the *stability condition*

$$\tfrac{1}{2}k\nu(S_2(2h))^2<1, \tag{6.11}$$

or more precisely

$$\tfrac{1}{2}k\nu(S_2(2h))^2\leqq 1-\theta, \tag{6.12}$$

for some θ, $0<\theta<1$. If we replace $S_2(2h)$ by its expression from (5.17), (6.11) becomes

$$\frac{k\nu}{h^2}<1. \tag{6.13}$$

If instead of (6.2)-(6.4) we consider the traditional explicit finite-difference scheme in V_h, we obtain the stability condition

$$\tfrac{1}{2}k\nu(S_2(h))^2<1,$$

i.e.,

$$\frac{k\nu}{h^2} < \frac{1}{4}.$$

In conclusion, the IMG *algorithm allows a time-step four times larger than the usual explicit finite-difference scheme* and yields a similar result with essentailly four times fewer computations. Of course the IMG algorithm as depicted in (6.3), (6.4) is not fully explicit since the determination of z_h^n from (6.5) is implicit. However, we must remember that the z_h^n are small increments and therefore their determination can be made in a very rudimentary way.

A more complete analysis of the time-discretized version of the IMG algorithm in the context of nonlinear equations will be performed elsewhere, but we thought that it would be useful to indicate here another partial justification of the IMG algorithm.

REFERENCES

[B] A. BRANDT, *Multigrid methods*, in Proc. International Congress of Mathematicians, Berkeley, CA, 1986.

[C] J. CÉA, *Approximation variationnelle des problèmes aux limites*, Ann. Inst. Fourier (Grenoble), 14 (1964), pp. 345-444.

[CFNT] P. CONSTANTIN, C. FOIAS, B. NICOLAENKO, AND R. TEMAM, *Integral Manifolds and Inertial Manifolds for Dissipative Partial Differential Equations*, Appl. Math. Sci. 70, Springer-Verlag, Berlin, New York, 1988.

[FMT] C. FOIAS, O. MANLEY, AND R. TEMAM, *Modeling of the interaction of small and large eddies in two-dimensional turbulent flows*, Math. Model. Numer. Anal., 22 (1988), pp. 93-114.

[FNST1] C. FOIAS, B. NICOLAENKO, G. SELL, AND R. TEMAM, *Variétés inertielles pour l'équation de Kuramoto-Sivashinski*, C.R. Acad. Sci. Paris Sér. I, 301 (1985), pp. 286-288.

[FNST2] ———, *Inertial manifolds for the Kuramoto-Sivashinsky equation and an estimate of their lowest dimension*, J. Math. Pures Appl., 67 (1988), pp. 197-226.

[FST1] C. FOIAS, G. SELL, AND R. TEMAM, *Variétés inertielles des equations différentielles dissipatives*, C.R. Acad. Sci. Paris Sér I, 301 (1985), pp. 139-142.

[FST2] ———, *Inertial manifolds for nonlinear evolutionary equations*, J. Differential Equations, 73 (1988), pp. 309-353.

[H] W. HACKBUSCH, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, New York, 1985.

[JRT] F. JAUBERTEAU, C. ROSIER, AND R. TEMAM, *The nonlinear Galerkin method in computational fluid dynamics*, Appl. Numer. Math., to appear.

[Mc] S. F. MCCORMICK, *Multigrid Methods*, Society for Industrial and Applied Mathematics, Philadelphia, 1987.

[MS] J. MALLET-PARET AND G. SELL, *Inertial manifolds for reaction diffusion equations in higher space dimensions*, J. Amer. Math. Soc., 1 (1988), pp. 805-866.

[MT1] M. MARION AND R. TEMAM, *Nonlinear Galerkin methods*, SIAM J. Numer. Anal., 26 (1989), pp. 1140-1158.

[MT2] ———, *Nonlinear Galerkin methods: the finite element case*, to appear.

[T1] R. TEMAM, *Sur l'approximation des équations de Navier-Stokes*, C.R. Acad. Sci. Paris Sér. A, 262 (1966), pp. 219-221.

[T2] ———, *Navier-Stokes Equations*, Third edition, North-Holland, Amsterdam, New York, 1984.

[T3] ———, *Navier-Stokes Equations and Nonlinear Functional Analysis*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 1983.

[T4] ———, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci. 68, Springer-Verlag, Berlin, New York, 1988.

[T5] ———, *Dynamical systems, turbulence and the numerical solution of the Navier-Stokes equations*, in Proc. Eleventh International Conference on Numerical Methods in Fluid Dynamics, D. L. Dwoyer and R. Voigt, eds., Lecture Notes in Physics, Springer-Verlag, Berlin, New York, 1989.

[T6] R. TEMAN, *Variétés inertielles approximatives pour les équations de Navier-Stokes bidimensionnelles*, C.R. Acad. Sci. Paris Sér. II, 306 (1988), pp. 399-402.

[T7] ———, *Attractors for the Navier-Stokes equations, localization and approximation*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., to appear.

[T8] ———, *Approximation of attractors and application to scientific computing*, Internat. J. Numer. Methods Engrg., to appear.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 179-204, January 1990
© 1990 Society for Industrial and Applied Mathematics
010

THE BIFURCATION OF HOMOCLINIC AND PERIODIC ORBITS FROM TWO HETEROCLINIC ORBITS*

S.-N. CHOW†, B. DENG‡, AND D. TERMAN§

Abstract. Conditions are found for a unique homoclinic or periodic orbit to bifurcate from a heteroclinic loop for autonomous ordinary differential equations. This leads to a codimension 2 unfolding of a heteroclinic loop. This approach, based on an idea developed by Šil'nikov, reduces the problem to the study of bifurcation equations. The result is applied to various types of traveling wave solutions of the FitzHugh-Nagumo equations with a cubic nonlinear term.

Key words. heteroclinic orbit, homoclinic orbit, periodic orbit, Šil'nikov solution, exponential expansion, strong λ-lemma, Lyapunov-Schmidt reduction, bifurcation equation

AMS(MOS) subject classifications. 34A34, 34C28, 34C99

1. Introduction. This paper is concerned with the creation of homoclinic and periodic orbits from a pair of heteroclinic orbits of differential equations of the form

$$\dot{x} = f(x, \alpha), \quad x \in \mathbb{R}^N, \quad \alpha \in \mathbb{R}^2. \tag{1.1}$$

We assume that $a, b \in \mathbb{R}^N$ are hyperbolic equilibria of (1.1) for all α. By a heteroclinic solution from a to b we mean a solution $\Gamma(t)$ of (1.1) that satisfies

$$\lim_{t \to -\infty} \Gamma(t) = a \quad \text{and} \quad \lim_{t \to +\infty} \Gamma(t) = b.$$

A homoclinic solution to a is a solution of (1.1) that satisfies

$$\lim_{|t| \to \infty} \Gamma(t) = a.$$

In this paper we demonstrate that a bifurcation of homoclinic solutions must take place at a value α_0 for which there exists a pair of heteroclinic solutions from a to b and from b to a, if certain generic conditions are satisfied. Moreover, if we assume that in a parameter space there are two curves c_{ab} and c_{ba} that cross transversely at α_0 and correspond to $a \to b$ and $b \to a$ heteroclinic solutions, respectively, then we show that there are two curves c_{aa} and c_{bb} in parameter space emanating from α_0 that correspond to homoclinic solutions. The curve c_{aa} will be tangent to c_{ab} at α_0, and c_{bb} will be tangent to c_{ba} at α_0. We also consider the existence of periodic solutions of (1.1). We prove that the curves c_{aa} and c_{bb} form the boundary of a sector Λ; a periodic solution of (1.1) exists for precisely those values of α in Λ.

The motivation of this work comes from the study of reaction-diffusion equations. These are equations of the form

$$U_t = DU_{xx} + F(U, \lambda) \tag{1.2}$$

* Received by the editors March 15, 1988; accepted for publication (in revised form) March 19, 1989.

† Department of Mathematics, Michigan State University, East Lansing, Michigan 48824. The research of this author was supported in part by the National Science Foundation under grant DMS 8401719 and by the Defense Advanced Research Projects Agency.

‡ Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska 68588.

§ Department of Mathematics, Ohio State University, Columbus, Ohio 43210. The research of this author was supported in part by the National Science Foundation under grant DMS 8401719 and DMS 8702693.

where $U \in \mathbb{R}^N$, D is a nonnegative diagonal matrix, and $\lambda \in \mathbb{R}$ is a parameter. A traveling wave solution of (1.2) is a solution of the form $U(x, t) = V(z)$, $z = x + \theta t$. That is, it corresponds to solutions that appear to be traveling with constant shape and velocity. We are interested in traveling wave solutions of (1.2) that connect two rest points of (1.2). Assume that A, $B \in \mathbb{R}^N$ satisfies $F(A, \lambda) = F(B, \lambda) = \mathcal{O}$ for all λ. Here, $\mathcal{O}$ is the origin in $\mathbb{R}^N$. We consider traveling wave solutions of (1.2) that satisfy

$$\lim_{z \to -\infty} V(z) = A \quad \text{and} \quad \lim_{z \to +\infty} V(z) = B.$$

Note that a traveling wave solution satisfies the following system of ordinary differential equations:

$$DV'' - \theta V' + F(V, \lambda) = 0.$$

If we let $V' = W$, then this is equivalent to the first-order system

$$V' = W, \qquad DW' = \theta W - F(V, \lambda) \tag{1.3}$$

together with the boundary conditions

$$\lim_{z \to -\infty} (V(z), W(z)) = (A, \mathcal{O}) \quad \text{and} \quad \lim_{z \to +\infty} (V(z), W(z)) = (B, \mathcal{O}).$$

Hence, the problem of proving the existence of a traveling wave solution of (1.2) reduces to finding a heteroclinic solution of (1.3). Note that the speed θ is a parameter in (1.3). Hence, (1.3) depends on two parameters and is a special case of (1.1).

Many systems possess a variety of traveling wave solutions. A given system may have, for a given value of parameters, traveling fronts, pulses, multiple pulses, and periodic solutions. If a given system does have many traveling wave solutions, then the existence of some of them (the traveling fronts, perhaps) may be easy to prove, while the existence of others (pulses, perhaps) may be more difficult to prove. Our results demonstrate that the more complicated waves can arise as bifurcations of the simpler waves. In § 6 we present an example to illustrate this point.

The proofs of our results are based on an idea of Šil'nikov [8], [11], [12]. It begins with a Poincaré return map on certain proper cross sections of the heteroclinic orbits. With Šil'nikov's change of variables for these Poincaré maps, the problem reduces to a two-parameter family of transcendental equations. The uniqueness of homoclinic and periodic orbits follows from an Implicit Function Theorem argument. The existence of homoclinic and periodic orbits is derived from certain bifurcation equations that arise from the transcendental equations.

A precise statement of our results is given in § 2. In § 3 we define Poincaré maps on certain proper cross sections and explain the idea of Šil'nikov's change of variables for these Poincaré maps. In § 4 we prove the uniqueness of the homoclinic and periodic orbits. In § 5 we derive the bifurcation equations for the existence of the homoclinic and periodic orbits, and then prove the main results. In § 6 we show how our results apply to the FitzHugh-Nagumo equations.

We point out that we can weaken considerably the generic hypothesis of the results presented in this paper and still conclude that there exists a bifurcation of homoclinic orbits. In [3] we prove such a result. This result is described at the end of § 2. Of course, under the weaker hypotheses we cannot expect to obtain the detailed description of the nature of the bifurcating orbits obtained in this paper.

2. Statement of the main result. In this section we will first introduce the main hypotheses (H1)-(H8), and then state our main theorem. We also state our main theorem in [3] for comparison.

Consider a system of ordinary differential equations with parameter α

$$\dot{z} = F(z, \alpha) \tag{2.1}$$

where $z = (z^{(1)}, \cdots, z^{(d)}) \in \mathbb{R}^d$ and $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^2$. We assume that (2.1) has two distinct hyperbolic equilibria a_1 and a_2 for all α. We also assume that when $\alpha = (0, 0)$, (2.1) has a heteroclinic orbit Γ_{12} from a_1 to a_2 and another heteroclinic orbit Γ_{21} from a_2 to a_1.

Let $C_i = C_i(\alpha) = DF(a_i, \alpha)$, $i = 1, 2$, where D is the differentiation operator with respect to z. Let $\sigma_i = \sigma_i(\alpha)$ be the spectrum of C_i, i.e.,

$$\sigma_i(\alpha) = \{\lambda \in \mathbb{C} \mid \lambda \text{ is an eigenvalue of } C_i\}. \tag{2.2}$$

Let $\sigma_i^+ = \sigma_i^+(\alpha)$ and $\sigma_i^- = \sigma_i^-(\alpha)$ be defined as follows:

$$\sigma_i^+(\alpha) = \sigma_i \cap \{\lambda \in \mathbb{C} \mid \operatorname{Re} \lambda > 0\} \tag{2.3}$$

$$\sigma_i^-(\alpha) = \sigma_i \cap \{\lambda \in \mathbb{C} \mid \operatorname{Re} \lambda < 0\}. \tag{2.4}$$

It is clear from the hyperbolicity of equilibria a_i, $i = 1, 2$ and the existence of heteroclinic orbits Γ_{12} and Γ_{21} that for $\alpha = (0, 0)$,

$$\sigma_i^+ \neq \varnothing, \qquad \sigma_i^- \neq \varnothing, \tag{2.5}$$

and

$$\sigma_i = \sigma_i^+ \cup \sigma_i^-. \tag{2.6}$$

Let $\mu_i = \mu_i(\alpha)$ with

$$\mu_i(\alpha) = \min_{\lambda \in \sigma_i^+} \operatorname{Re} \lambda. \tag{2.7}$$

Then (2.4) and (2.5) imply

$$\mu_i > 0. \tag{2.8}$$

First, we consider two conditions on the eigenvalues that are not required in [3].

(H1) μ_i is a simple real eigenvalue for C_i for $i = 1, 2$ and

$$\mu_i < \min_{\lambda \in \sigma_i^+ - \{\mu_i\}} \operatorname{Re} \lambda, \qquad i = 1, 2, \tag{2.9}$$

(H2) $$\mu_i < -\max_{\lambda \in \sigma_i^-} \operatorname{Re} \lambda, \qquad i = 1, 2.$$

For many problems of homoclinic bifurcations, hypotheses (H1) and (H2) play a crucial role in determining the bifurcation structure. For example, it is well known that if the first eigenvalue of σ^+ is a pair of complex eigenvalues rather than a simple real one as in (H1), then the system has an invariant set carrying the Bernoulli shift (see Šil'nikov [18]). Also, if the largest eigenvalue of σ^- is real and simple and if its absolute value is equal to the first eigenvalue of σ^+ (this will violate (H2)), then in any small neighborhood of the homoclinic orbit the periodic orbits are no longer unique for small $\alpha = 0$. Furthermore, double-periodic and double-homoclinic orbits may occur in the latter case (see Yanagida [16] and Chow, Deng, and Fiedler [17]). However, as we will show, if (H1) and (H2) are satisfied in addition to other hypotheses, then the bifurcations of homoclinic and periodic orbits from the heteroclinic loop $\Gamma_{12} \cup \Gamma_{21}$ are unique.

Let $m_i > 0$ and $n_i > 0$ be the dimensions of the stable and unstable manifolds of a_i, respectively. As in [3], we assume the following hypothesis.

(H3) $m_1 = m_2 = m$ and $n_1 = n_2 = n$.

Because of (H3) it is well known (see Hale and Lin [6] and Palmer [9]) that the continuation of heteroclinic orbits $\Gamma_{12}(\Gamma_{21})$ occurs generically in a codimension-1 submanifold in the space of vector fields. This implies that (H4)-(H6) are all generic.

(H4) Γ_{12} and Γ_{21} are in general position:

$$\text{codim span}\{T_pW_1^u, T_pW_2^s\} = 1, \qquad p \in \Gamma_{12},$$

$$\text{codim span}\{T_pW_2^u, T_pW_1^s\} = 1, \qquad p \in \Gamma_{21},$$

where W_i^u and W_i^s denote the stable and unstable manifolds of a_i for $i = 1, 2$ at $\alpha = (0, 0)$, and T_pW denotes the tangent space of a given manifold W at basepoint p.

We also assume that there are two smooth curves intersecting transversally in the parameter spaces that correspond to the smooth branches of heteroclinic orbits from a_1 to a_2 and from a_2 to a_1, respectively. Thus, up to a change of coordinates in parameter space, we assume the following hypothesis.

(H5) Equation (2.1) has a smooth branch of heteroclinic orbits Γ_{α_1} from a_1 to a_2 with $(\alpha_1, 0) \in \mathbb{R}^2$ and $\Gamma_0 = \Gamma_{12}$. Equation (2.1) has a smooth branch of heteroclinic orbit Γ_{α_2} from a_2 to a_1 with $(0, \alpha_2) \in \mathbb{R}^2$ and $\Gamma_0 = \Gamma_{21}$.

The next hypothesis (H6) is related to the transverse crossing of the stable and unstable manifolds as parameters vary. To be more precise, let $W_i^s(\alpha)$ and $W_i^u(\alpha)$ denote the stable and unstable manifolds of a_i, $i = 1, 2$ for (2.1) at α. Clearly, $W_i^s(0, 0) = W_i^s$ and $W_i^u(0, 0) = W_i^u$, where W_i^s and W_i^u are as in (H4). Let Σ_1 be an arbitrary and small $(d-1)$-dimensional cross section such that Σ_1 intersects Γ_{12} at exactly one point, and Σ_1 is transverse to the flow of (2.1) for $\alpha = (0, 0)$. Let $M_1^u = M_1^u(\alpha)$ and $M_2^s = M_2^s(\alpha)$ be connected components of $W_1^u(\alpha) \cap \Sigma_1$ and $W_2^s(\alpha) \cap \Sigma_1$, respectively, satisfying that M_1^u and M_2^s vary continuously with α near $\alpha = 0$ and

$$M_1^u(0) \cap M_2^s(0) = \Gamma_{12} \cap \Sigma_1. \tag{2.10}$$

Let $d_1 = d_1(\alpha_1, \alpha_2)$ be the distance between M_1^u and M_2^s defined by

$$d_1(\alpha_1, \alpha_2) = \inf |z_1 - z_2| \qquad \text{with } z_1 \in M_1^u \quad \text{and} \quad z_2 \in M_2^s. \tag{2.11}$$

Similarly, we can choose a proper cross section Σ_2 transverse to Γ_{21} and their distance $d_2 = d_2(\alpha_1, \alpha_2)$. It is easy to see that hypothesis (H5) implies

$$d_1(\alpha_1, 0) = 0. \tag{2.12}$$

Similarly, we have

$$d_2(0, \alpha_2) = 0. \tag{2.13}$$

We assume hypothesis (H6).

(H6) W_1^u and W_2^s, and W_2^u and W_1^s intersect transversally in the following sense:

$$\lim_{\alpha_2 \to 0} \frac{d_1(0, \alpha_2)}{|\alpha_2|} \neq 0 \quad \text{and} \quad \lim_{\alpha_1 \to 0} \frac{d_2(\alpha_1, 0)}{|\alpha_1|} \neq 0.$$

Note that (H6) implies $d_1(0, \alpha_2)$ has the same order as $|\alpha_2|$. This is certainly a generic assumption. The conditions here appear weaker than the usual conditions on the transverse crossing of stable and unstable manifolds. In § 5 we will see that together with the other hypotheses, (H6) does imply the transverse crossing in the usual sense. Indeed, the Melnikov functions in $\mathbb{R}^2$ are precisely the distances d_1 and d_2. Also, from (2.11) we know that d_1 depends on the choice of cross section Σ_1. However, it is not difficult to see that because the map induced by the flow of (2.1) from one cross section to another is a diffeomorphism, assumption (H6) is actually independent of the choice of cross sections.

It is clear that hypotheses (H4)-(H6) are concerned with the global structure of the stable and unstable manifolds. The following two hypotheses, (H7) and (H8), however, are concerned with both the global and local structure of the unstable manifolds near the equilibria. It is shown in Deng [5] that they are generic assumptions.

(H7) Γ_{12} is tangent to an eigenvector of $C_1(0)$ for the eigenvalue $\mu_1(0)$ as $t \to -\infty$; Γ_{21} is tangent to an eigenvector of $C_2(0)$ for the eigenvalue $\mu_2(0)$ as $t \to -\infty$.

Hypothesis (H8) is concerned with the inclination behavior of the global unstable manifold of one equilibrium near the other equilibrium. This has to do with the local strong unstable manifold W_i^{uu} of a_i that is $(n-1)$-dimensional and is tangent at a_i to the linear subspace spanned by the eigenvectors of C_i corresponding to the eigenvalues $\lambda \in \sigma_i^+ - \{\mu_i\}$. To be more precise, let (x, y) be local coordinates of points in a sufficiently small neighborhood U_i of a_i such that $x = 0$ and $y = 0$ correspond to the subspaces spanned by eigenvectors of $\lambda \in \sigma_i^+$ and $\lambda \in \sigma_i^-$, respectively. In particular, choose the $y^{(1)}$-axis as the direction of an eigenvector for the principal eigenvalue μ_i of C_i. Then W_i^{uu} can be expressed as the graph of a smooth function h^{uu} of the variable $\hat{y} = (y^{(2)}, \cdots, y^{(n)})$ that parameterizes the strong unstable eigenvector subspace of eigenvalues $\lambda \in \sigma_i^+ - \{\mu_i\}$.

DEFINITION. An $(n-1)$-dimensional smooth manifold D^{n-1} having nonempty intersection with the stable manifold W_i^s satisfies the *strong inclination property* if for every $\varepsilon > 0$ there is a $T(\varepsilon) > 0$ such that for every $t \geqq T(\varepsilon)$ the connected component of the image D_t^{n-1} in U_i under the time t mapping of the solutions with initial data from D^{n-1} can be expressed as the graph of a smooth function h_t of the same argument $\hat{y}$ as h^{uu} satisfying $\|h_t - h^{uu}\|_{C^1} < \varepsilon$, where $\|\cdot\|_{C^1}$ denotes the usual C^1 norm.

It is shown in Deng [5] that this strong inclination property holds true for a generic family of such D^{n-1}. The corresponding result is referred to as the strong λ-lemma for D^{n-1} in [5]. Now, let $M_1^u(0)$ and $M_2^u(0)$ be as in (2.10) and (H6). They are $(n-1)$-dimensional and intersect $W_2^s(0)$ and $W_1^s(0)$ at a single point, respectively. We assume hypothesis (H8).

(H8) $M_1^u(0)$ and $M_2^u(0)$ satisfy the strong inclination property (cf. Fig. 2.1).

Note that by Deng [5], (H7) is equivalent to $\Gamma_{12} \cap W_1^{uu} = \varnothing$ and $\Gamma_{21} \cap W_2^{uu} = \varnothing$. Also, due to the group property of the flow, it is not difficult to see that (H8) is independent of the choice of the cross section Σ_1 and Σ_2 in the definitions of M_1^u and M_2^u. Together with (H4), hypothesis (H8) will enable us to choose a one-dimensional subspace complementary to the $(d-2)$-dimensional subspace of span $\{T_p W_2^s(0), T_p M_1^u(0)\}$ in $T_p \Sigma_1 = \mathbb{R}^{d-1}$ at $p = W_2^s(0) \cap M_1^u(0) = \Gamma_{12} \cap \Sigma_1$. We will see that if Σ_1 is sufficiently close to the equilibrium a_2, then this complementary subspace can be chosen approximately to be the eigenvector subspace of the principal positive eigenvalue $\mu_2(0)$ for $C_2(0)$ provided (H8) is satisfied. All this will be done in §§ 4 and 5. As shown by the following main result, a system satisfying the eigenvalue conditions

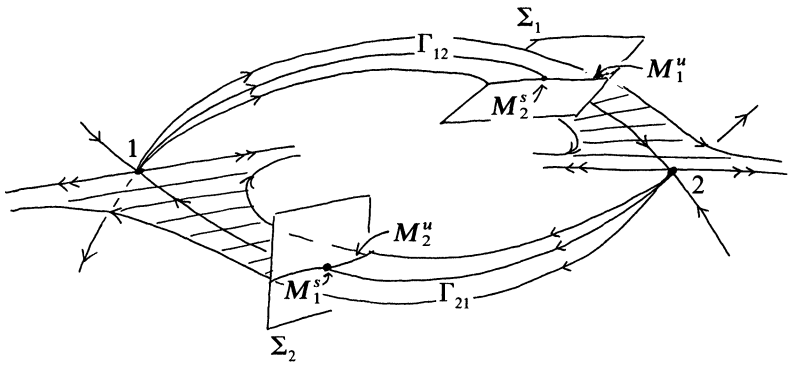


FIG. 2.1

(H1)-(H3) as well as the strong inclination properties (H7) and (H8) is analogous to the classical system considered by Šil'nikov in [12] which has a homoclinic orbit and satisfies conditions (H1), (H2), (H7), and (H8).

THEOREM 2.1. *Suppose F is C^k with $k \geqq 4$ and* (H1)-(H8) *are satisfied. Then there exists a small neighborhood $N \subset \mathbb{R}^d$ of $\Gamma_{12} \cup \Gamma_{21} \cup \{a_1, a_2\}$, a small neighborhood $V \subset R^2$ of $\alpha = (0, 0)$, and a nonsingular change of parameters $\varepsilon = c(\alpha)$, $\varepsilon \in E = c(V)$, such that the following holds true.*

(a) *There exist smooth functions $\varepsilon_1 = k_1(\varepsilon_2)$ defined for $\varepsilon_2 > 0$, $(\varepsilon_1, \varepsilon_2) \in E$, and $\varepsilon_2 = k_2(\varepsilon_1)$ defined for $\varepsilon_1 > 0$, $(\varepsilon_1, \varepsilon_2) \in E$ such that if*

$$\mathscr{C}_1 = \{\varepsilon \in E \mid \varepsilon_1 = k_1(\varepsilon_2),\ \varepsilon_2 > 0\}$$

and

$$\mathscr{C}_2 = \{\varepsilon \in E \mid \varepsilon_2 = k_2(\varepsilon_1),\ \varepsilon_1 > 0\},$$

then, for $i = 1, 2$, (2.1) *has a homoclinic orbit $\gamma \subset N$ to a_i with parameter α if and only if $\varepsilon = c(\alpha) \in \mathscr{C}_i$. Moreover, for each α with $c(\alpha) \in \mathscr{C}_i$, γ is unique in N. Furthermore,*

$$(\varepsilon_1, 0) = c((\alpha_1, 0)) \quad \text{and} \quad (0, \varepsilon_2) = c((0, \alpha_2)), \tag{2.14}$$

$$\lim_{\varepsilon_2 \to 0^+} k_1(\varepsilon_2) = \lim_{\varepsilon_1 \to 0^+} k_2(\varepsilon_1) = 0, \tag{2.15}$$

$$\lim_{\varepsilon_2 \to 0^+} \frac{dk_1}{d\varepsilon_2}(\varepsilon_2) = \lim_{\varepsilon_1 \to 0^+} \frac{dk_2}{d\varepsilon_1}(\varepsilon_1) = 0. \tag{2.16}$$

(b) *Let*

$$\Lambda = \{\varepsilon = (\varepsilon_1, \varepsilon_2) \in E \mid \text{either } \varepsilon_1 > 0 \text{ or } \varepsilon_2 > 0, \tag{2.17}$$
$$\varepsilon_1 > k_1(\varepsilon_2) \text{ if } \varepsilon_2 > 0, \text{ and } \varepsilon_2 > k_2(\varepsilon_1) \text{ if } \varepsilon_1 > 0\}.$$

Then, (2.1) *has a periodic orbit $\gamma \subset N$ at parameter α if and only if $c(\alpha) \in \Lambda$. Finally, for each parameter α with $c(\alpha) \in \Lambda$, γ is the unique periodic orbit in N.*

Figure 2.2 is the bifurcation diagram for Theorem 2.1.

As we mentioned in the Introduction, we can weaken the hypotheses of Theorem 2.1 and still prove the existence of bifurcating homoclinic orbits. In [3] we prove the following result.

THEOREM 2.2. *Suppose F is C^2 and hypotheses* (H3)-(H6) *are satisfied. Then there is a continuous map $\kappa : (0, 1] \to \mathbb{R}^2$ such that for each $s \in (0, 1]$,* (2.1) *has a homoclinic orbit to a_1 for $\alpha = \kappa(s)$. If $s_1 \neq s_2$, then the corresponding connections are not the same. Moreover, $\lim_{s \to 0^+} \kappa(s) = (0, 0)$. Finally, the same conclusion applies to a_2.*

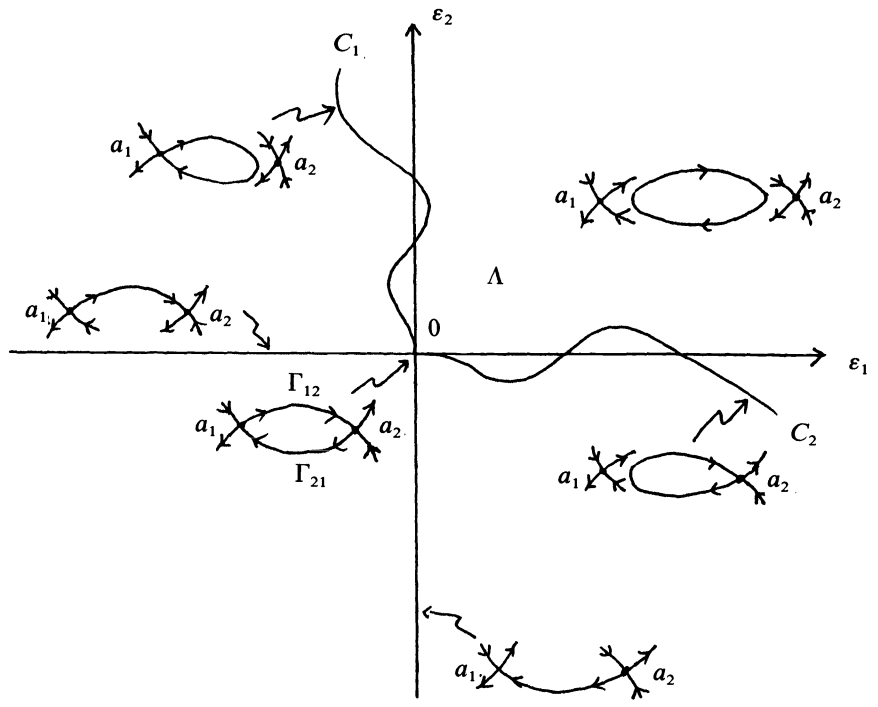


FIG. 2.2

Remark 2.3. The eigenvalue conditions (H1) and (H2) and the strong inclination properties (H7) and (H8) are required for the implicit function principle argument in this paper, which results in the uniqueness of the homoclinic and periodic orbits bifurcating from the loop $\Gamma_{12} \cup \Gamma_{21}$. These conditions are not required in Theorem 2.2. Theorem 2.2 asserts the bifurcation of homoclinic orbits must take place even though uniqueness is not claimed. The importance of Theorem 2.2 as well as its topological approach presented in [3] is to help us understand better the structure of vector fields near a codimension-2 bifurcation point at which a heteroclinic loop $\Gamma_{12} \cup \Gamma_{21}$ takes place. It tells us that this bifurcation point is precisely located at the intersection of the closures of two codimension-1 bifurcation branches on which the homoclinic bifurcations take place. Similar topological structures near a vector field that represents higher than codimension-2 bifurcation points at which a heteroclinic loop takes place should also be expected. As suggested by Theorem 2.2, this has to do with a further relaxation on the condition of equal dimensionality $\dim W_1^s = \dim W_2^s = m$ and $\dim W_1^u = \dim W_2^u = n$ required by both Theorems 2.1 and 2.2. Also note that the vector fields in Theorem 2.2 are assumed to be only C^2.

3. Poincaré maps and Šil'nikov's change of variables. A natural approach to this bifurcation problem is through the study of Poincaré maps for the loop $\Gamma_{12} \cup \Gamma_{21} \cup \{a_1, a_2\}$. To do this we need a result by Deng [5].

THEOREM 3.1. *Suppose that the system of ordinary differential equations with parameter* α,

$$\begin{aligned} \dot{x} &= A(\alpha)x + f(x, y, \alpha), \\ \dot{y} &= B(\alpha)y + g(x, y, \alpha), \end{aligned} \qquad x \in \mathbb{R}^m, \quad y \in \mathbb{R}^n, \quad \alpha \in \mathbb{R}^2 \tag{3.1}$$

satisfies the following hypotheses:

(H9) *The matrix functions* $A = A(\alpha) \in \mathbb{R}^{m\times n}$, $B = B(\alpha) \in \mathbb{R}^{n\times n}$ *and the vector functions* $f = f(x, y, \alpha)$, $g = g(x, y, \alpha)$ *are* C^k *with* $k \geqq 3$,

(H10) $f(0, y, \alpha) = 0, \qquad g(x, 0, \alpha) = 0,$

$Df(0, 0, \alpha) = 0, \qquad Dg(0, 0, \alpha) = 0$

where D *is the differentiation operator with respect to* $z = (x, y)$,

(H11) B *has the following form*:

$$B = \begin{bmatrix} \mu(\alpha) & 0 \\ 0 & B_1(\alpha) \end{bmatrix} \tag{3.2}$$

and satisfies

$$0 < \mu(\alpha) < \min_{\lambda \in \sigma(B_1(\alpha)) - \{\mu(\alpha)\}} \operatorname{Re} \lambda, \tag{3.3}$$

$$0 < \mu(\alpha) < -\min_{\lambda \in \sigma(A(\alpha))} \operatorname{Re} \lambda. \tag{3.4}$$

Then there exist sufficiently small constants $\delta_0 > 0$ *and* α_0 *such that for every* $s \geqq 0$, $|x_0| \leqq \delta_0$, $|y_1| \leqq \delta_0$, *and* $|\alpha| \leqq \alpha_0$, (3.1) *has a unique solution* $x(t) = x(t; s, x_0, y_1, \alpha)$, $y(t) = y(t; s, x_0, y_1, \alpha)$ *satisfying*

$$x(0) = x_0, \qquad y(s) = y_1, \tag{3.5}$$

$$|x(t)| \leqq 2\delta_0, \quad |y(t)| \leqq 2\delta_0, \quad 0 \leqq t \leqq t_0 \tag{3.6}$$

where $t_0 > s$ *is some constant depending on* δ, s, x_0, *and* y_1. *Furthermore, as functions of* t, s, x_0, y_1 *and* α *the solution* $(x, y)(t; s, x_0, y_1, \alpha)$ *is* C^k *and* C^{k+1} *in* t. *Also, there exist constants* $K_0 \geqq 1$ *and* $\bar{\nu} > 0$, *which depend only on* $\mu(\alpha)$ *and* $\lambda(\alpha)$, *and a* C^{k-2} *function* $\varphi = \varphi(x_0, y_1, \alpha) \in \mathbb{R}^n$ *such that*

$$\sum_{j=0}^{k-1} |D^j x(s; s, x_0, y_1, \alpha)| \leqq K_0 e^{-\lambda(\alpha)s}, \tag{3.7}$$

$$\sum_{j=0}^{k-2} |D^j (e^{\mu(\alpha)s} y(0; s, x_0, y_1, \alpha) - \varphi(x_0, y_1, \alpha))| \leqq K_0 e^{-\nu s} \tag{3.8}$$

where D^j *is the differentiation operator of order* j *in* (s, x_0, y_1, α). *Moreover, the function* φ *satisfies*

$$D\varphi(0, 0, \alpha) = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{n\times d} \tag{3.9}$$

where D *is the first-order differentiation operator in* (x_0, y_1). *Finally, the local strong unstable manifold* W^{uu}_{loc} *can be expressed as follows*:

$$\{(0, y) \mid \varphi^{(1)}(0, y, \alpha) = 0, |y| \leqq \delta_0\} = W^{uu}_{\text{loc}} \tag{3.10}$$

where $\varphi = (\varphi^{(1)}, \cdots, \varphi^{(n)})$.

Remark 3.2. (a) Note that if $s = 0$, then $x(t) = x(t; 0, x_0, y_1, \alpha)$, $y(t) = y(t; 0, x_0, y_1, \alpha)$ is the unique solution of the initial value problem $x(0) = x_0$, $y(0) = y_0$ with $y_0 = y_1$ for (3.1). This kind of boundary value problem was first introduced by Šil'nikov [11] for the study of the structure of flows near a homoclinic orbit.

(b) We will not need the exponential estimates (3.7) and (3.8) in Theorem 3.1 in this section. However, they are useful in later sections. In particular, the estimate (3.8) is extremely important and is equivalent to writing $y(0; s, x_0, y_1, \alpha) = \varphi(x_0, y_1, \alpha)e^{-\mu(\alpha)s} + R(s, x_0, y_1, \alpha)$ with a C^{k-2} function R all of whose derivatives up to order $k-2$ are exponentially bounded by $K_0 e^{-(\mu(\alpha)+\bar{\nu})s}$. This decomposition of $y(0)$ is referred to as an exponential expansion in [5].

(c) The properties (3.9) and (3.10) on the coefficient function φ for the exponential expansion will be used later to construct a one-dimensional subspace complementary to the span$\{T_p W_2^s(0), T_p M_1^u(0)\}$ in $T_p\Sigma_1 = \mathbb{R}^{d-1}$ at $p = W_2^s(0) \cap M_1^u(0) = \Gamma_{12} \cap \Sigma_1$ mentioned earlier when (H8) on the strong inclination property for $M_1^u(0)$ was introduced.

To apply Theorem 3.1, we first observe that by a smooth change of variables in a small neighborhood U_i of a_i, (2.1) is C^k conjugate in U_i to equations in the following form:

$$\begin{aligned} \dot{x} &= A_i(\alpha)x + f_i(x, y, \alpha), \\ \dot{y} &= B_i(\alpha)y + g_i(x, y, \alpha), \end{aligned} \qquad x \in \mathbb{R}^m, \quad y \in \mathbb{R}^n, \quad \alpha \in \mathbb{R}^2. \tag{3.11}$$

Here A_i, B_i, f_i, g_i and C^{k-1}, and f_i and g_i satisfy (H10) for $i = 1, 2$. Furthermore, (H1) and (H2) imply that A_i and B_i satisfy (H11). Thus, Theorem 3.1 is applicable to (3.11) for $i = 1$ and 2. It is not difficult to see that we may choose a single δ_0 from Theorem 3.1 such that the conclusions hold true for (3.11) for $i = 1$ and 2. We will use this δ_0 to define several cross sections to the loop $\Gamma_{12} \cup \Gamma_{21} \cup \{a_1, a_2\}$ and to the maps along them.

Since the discussion in the following few paragraphs applies to both a_1 and a_2, we will treat only a_1 and, therefore, suppress all the subscripts of a. Also, we will suppress the parameter α if doing so does not cause confusion.

First, we specify some notation used in this paper. Let $p \in W_{\text{loc}}^s \cap \{\Gamma_{12} \cup \Gamma_{21}\}$ and $q \in W_{\text{loc}}^u \cap \{\Gamma_{12} \cup \Gamma_{21}\}$. Write

$$p = (\bar{x}_0, 0), \qquad q = (0, \bar{y}_1). \tag{3.12}$$

By (H7), we may choose $\bar{y}_1 = (\bar{y}_1^{(1)}, \cdots, \bar{y}_1^{(n)})$ to satisfy

$$\bar{y}_1^{(1)} = \delta_0. \tag{3.13}$$

For simplicity, we assume that

$$\bar{x}_0^{(1)} = \delta_0 \tag{3.14}$$

where $\bar{x}_0 = (\bar{x}_0^{(1)}, \cdots, \bar{x}_0^{(m)})$. We remark that (3.14) will not be used in any proof and therefore is not essential. Let $\delta_0 > \delta_1 > 0$ and $\delta_0 > \delta_2 > 0$ be arbitrary small constants, and

$$\Sigma^s(\delta_1) = \{(x, y) | x^{(1)} = \delta_0, |x - \bar{x}_0| < \delta_1, |y| < \delta_1\}, \tag{3.15}$$

$$\Sigma^u(\delta_2) = \{(x, y) | y^{(1)} = \delta_0, |x| < \delta_2, |y - \bar{y}_0| < \delta_2\}. \tag{3.16}$$

Note that for sufficiently small δ_2, (H1) implies that $\Sigma^u = \Sigma^u(\delta_2)$ is transverse to the flow of (2.1). For simplicity, we assume, without loss of generality, that $\Sigma^s = \Sigma^s(\delta_1)$ is also transverse to the flow of (3.1) for small δ_1. Again, we emphasize that the forms (3.14) and (3.15) are merely for simplicity in our discussion and will not be used in our proof. Let $z(t) = z(t; z_0)$ denote the solution of (3.11) with initial data $z(0) = z_0$, and let $z = (x, y)$.

Next, we define a local map near the equilibrium by the flow. Define

$$\begin{aligned} \sigma^s = \{(x_0, y_0) \in \Sigma^s \,|\, &\exists s = s(x_0, y_0) \text{ such that } z(t; z_0) \notin \Sigma^u \\ &\text{for } 0 \leqq t < s \text{ and } z(s; z_0) \in \Sigma^u, \text{ where } z_0 = (x_0, y_0)\}. \end{aligned} \tag{3.17}$$

It is not difficult to see that since Σ^s and Σ^u are transverse to the flow of (3.11), δ_1 and δ_2 can be chosen so small that the function $s: \sigma^s \to \mathbb{R}$, which represents the first time needed for the trajectory starting from σ^s to reach Σ^u, is well defined and continuously differentiable. Define

$$\pi: \sigma^s \to \Sigma^u \tag{3.18}$$

$$(x_0, y_0) \to z(s; z_0) \text{ with } s = s(x_0, y_0) \text{ and } z_0 = (x_0, y_0).$$

Obviously, π is a diffeomorphism onto its image. Let

$$\sigma^u = \pi(\sigma^s) \tag{3.19}$$

(see Fig. 3.1).

Applying the same arguments (3.12), (3.15)-(3.19) to (3.11) for $i = 1$ and 2, we obtain the following points, maps, etc.: p_i, q_i, δ_{i1}, δ_{i2}, Σ_i^s, Σ_i^u, σ_i^s, σ_i^u, and $\pi_i: \sigma_i^s \to \sigma_i^u$, $i = 1, 2$.

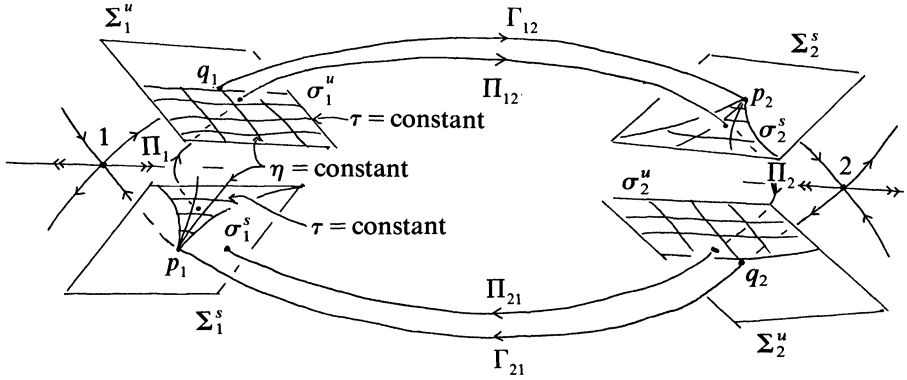


FIG. 3.1

Next, we define a global map near Γ_{12} by following the flow. Since q_1 and p_2 are on Γ_{12}, there exists a unique $t_1 > 0$ such that $z(t; q_1) \notin \Sigma_2^s$ for $0 \leqq t < t_1$ and $z(t_1; q_1) = p_2$. Hence, by the Implicit Function Theorem, for every $(x_0, y_0) \in \Sigma_1^u$ sufficiently close to q_1, there exists a unique time $t_2 = t_2(x_0, y_0) > 0$, such that $z(t; z_0) \notin \Sigma_2^s$ for all $0 \leqq t < t_2$, $z_0 = (x_0, y_0)$, and $z(t_2; z_0) \in \Sigma_2^s$, that is, the first time for the trajectory to hit Σ_2^s. It is easy to see that we can choose δ_{12} sufficiently small so that $t_2(x_0, y_0)$ is smoothly defined for all $(x_0, y_0) \in \Sigma_1^u$ and remains close to the constant t_1. Define

$$\pi_{12}: \Sigma_1^u \to \Sigma_2^s \tag{3.20}$$

$$(x_0, y_0) \to z(t_2, z_0) \text{ with } t_2 = t_2(x_0, y_0) \text{ and } z_0 = (x_0, y_0).$$

Obviously, π_{12} is a diffeomorphism. Similarly, for a sufficiently small δ_{22}, we can define $\pi_{21}: \Sigma_2^u \to \Sigma_1^s$.

Note that if

$$\pi_{12}(\sigma_1^u) \subset \sigma_2^s \tag{3.21}$$

then, a Poincaré return map $\pi = \pi_{21} \cdot \pi_{12} \cdot \pi_{12} \cdot \pi_1$: $\sigma_1^s \to \Sigma_1^s$ is well defined. Since we can always expect that (3.21) holds true for certain subset $\tilde{\sigma}_1^u \subset \sigma_1^u$, a map

$$\pi_{11} = \pi_{21} \cdot \pi_2 \cdot \pi_{12} \cdot \pi_1: \tilde{\sigma}_1^s \to \Sigma_1^s \tag{3.22}$$

is always well defined. π_{11} is called a Poincaré map. Similarly, we may obtain the other Poincaré map π_{22} (cf. Fig. 3.1).

Now, the following discussion is devoted to the relation between the periodic points of π_{11} and the reduced equivalent equations. Naturally, a periodic orbit near $\Gamma_{12} \cup \Gamma_{21}$ corresponds to a periodic point of either π_{11} or π_{22}. Intuitively, the structure of $\tilde{\sigma}_1^s$ is more complicated than that of σ_1^s. Thus, instead of studying the fixed points of the Poincaré map π_{11} on $\tilde{\sigma}_1^s$, we consider the following equations:

$$\begin{aligned} &\pi_{12} \cdot \pi_1(z_1) = z_2, \\ &\pi_{21} \cdot \pi_2(z_2) = z_1, \end{aligned} \qquad z_1 \in \sigma_1^s, \quad z_2 \in \sigma_2^s. \tag{3.23}$$

It is obvious that the existence of fixed points for π_{11} is equivalent to the existence of solutions z_1 and z_2 to (3.23). However, when we examine (3.17) and (3.18) for π_1 (equivalently, π_2), we find several difficulties. First, since $p_1 \in \Gamma_{21}$ is on the local stable manifold of a_1, by (3.17) p_1 must not be in σ_1^s. Second, due to the hyperbolicity of the equilibrium a_1, the time $s_1 = s_1(x_0, y_0)$ needed for a point on the orbit to travel from σ_1^s to σ_1^u approaches infinity as $(x_0, y_0) \to p_2$. Moreover, the definitions for π_1 and π_2 reveal few properties with which we can analyze (3.23). Usually, if we can find a local change of variables near each a_1 and a_2 such that under the new variables the nonlinear equation (3.1) becomes a system of linear equations in U_1 and U_2 then these problems will be dramatically simplified. This so-called C^1-linearization approach is indeed quite common in literature. Our hypotheses (H1) and (H2) on the eigenvalues are not sufficient to obtain such a C^1-linearization in general. The approach to overcome these difficulties is through a change of variables only on σ_1^s and σ_2^s. This idea is due to Šil'nikov.

Let Δ_1 be a $(d-1)$-dimensional open set. Let $\rho_1 \colon \Delta_1 \to \sigma_1^s$ be a diffeomorphism onto its image. Then ρ_1 is called a *change of variables* in σ_1^s. Let $\rho_{11} = \pi_1 \cdot \rho_1 \colon \Delta_1 \to \sigma_i^u$. Similarly, we may have a change of variables ρ_2, in σ_2^s. The purpose of such changes of variables is to make the local maps ρ_{11} and ρ_{22} tractable. Before we introduce the Šil'nikov changes of variables, we note the following simple facts.

It is easy to see that to find solutions z_1 and z_2 to (3.23) it suffices to find solutions to the following equations:

$$\begin{aligned} &\pi_{12} \cdot \rho_{11}(\zeta_1) = \rho_2(\zeta_2), \\ &\pi_{21} \cdot \rho_{22}(\zeta_2) = \rho_1(\zeta_1), \end{aligned} \qquad \zeta_1 \in \Delta_1, \quad \zeta_2 \in \Delta_2. \tag{3.24}$$

However, the existence of solutions to (3.24) is only a sufficient condition for the existence of solutions to (3.23). But we will see that by restricting periodic orbits of (2.1) to a sufficiently small neighborhood N of $\Gamma_{12} \cup \Gamma_{21} \cup \{a_i, a_2\}$, we can guarantee that the condition for the existence of solutions to (3.24) in some subsets $\hat{\Delta}_1 \subset \Delta_1$ and $\hat{\Delta}_2 \subset \Delta_2$ is both sufficient and necessary for the existence of certain types of periodic orbits in N. This is to be explained as follows.

In this paper, a small neighborhood N of $\Gamma_{12} \cup \Gamma_{21} \cup \{a_1, a_2\}$ satisfies

$$N \cap L_i^s \subset \Sigma_i^s, \qquad i = 1, 2 \tag{3.25}$$

where $L_i^s = \{(x, y) \in U_i \mid x^{(1)} = \delta_0, |x| \leqq \delta_0, |y| \leqq \delta_0\}$. Let

$$\hat{\Sigma}_i^s = N \cap \Sigma_i^s, \qquad i = 1, 2, \tag{3.26}$$

$$\hat{\sigma}_i^s = \hat{\Sigma}_i^s \cap \sigma_i^s, \qquad i = 1, 2. \tag{3.27}$$

Let γ be a periodic (homoclinic) orbit in N; then γ is called a *K-periodic (homoclinic) orbit* if the number of points in the set $\gamma \cap \hat{\sigma}_1^s (\gamma \cap \operatorname{cl} \hat{\sigma}_1^s)$ is exactly K. Certainly, for a K-periodic (homoclinic) orbit γ in N the number of points in $\gamma \cap \hat{\sigma}_2^s (\gamma \cap \operatorname{cl} \hat{\sigma}_2^s)$ is also equal to K. Moreover, if N is so small that

$$\hat{\sigma}_i^s \subset \rho_i(\Delta_i), \qquad i = 1, 2, \tag{3.28}$$

then there exists a K-periodic orbit γ in N, say $K=1$, if and only if (3.24) has solutions $\zeta_1 \in \hat{\Delta}_1$ and $\zeta_2 \in \hat{\Delta}_2$ where

$$\hat{\Delta}_i = \rho_i^{-1}(\hat{\sigma}_i^s), \qquad i=1,2. \tag{3.29}$$

We will put these observations into a lemma. Before doing so, we note that all definitions for maps and sets as above extend immediately to the perturbed system of (2.1) with a small parameter α. Thus, from now on we allow the arguments of π_1, π_2, π_{12}, ρ_2, ρ_{11}, ρ_{22}, etc. to include the parameter α. But we write σ_1^s instead of $\sigma_1^s(\alpha)$, and so on.

LEMMA 3.3. *Let σ_i^s, σ_i^u, π_i, $i=1,2$ be defined as in* (3.17)-(3.19). *Let π_{12}, π_{21} be defined as in* (3.20). *Let ρ_i, $i=1,2$ be a smooth map satisfying that for every α, $\rho_i(\cdot,\alpha):\Delta_i(\alpha)\to\sigma_i^s(\alpha)$ is a diffeomorphism onto its images. Let*

$$\rho_{ii}(\cdot,\alpha)=\pi_i(\rho_i(\cdot,\alpha),\alpha):\Delta_i(\alpha)\to\sigma_i^u(\alpha), \qquad i=1,2. \tag{3.30}$$

Then the following statements hold true:

(a) *For each small α there exist a small neighborhood N of $\Gamma_{12}\cup\Gamma_{21}\cup\{a_1,a_2\}$ and a K-periodic orbit γ in N if and only if the equations*

$$\begin{aligned}&\pi_{12}(\rho_{11}(\zeta_i^{(j)},\alpha),\alpha)=\rho_2(\zeta_2^{(j)},\alpha),\\ &\pi_{21}(\rho_{22}(\zeta_2^{(j)},\alpha),\alpha)=\rho_1(\zeta_1^{(j+1)},\alpha),\end{aligned} \qquad j=0,1,2,\cdots(\mathrm{mod}\ K) \tag{3.31}$$

have a solution

$$\zeta_i^{(j)}\in\hat{\Delta}_i(\alpha), \qquad i=1,2, \quad j=0,1,2,\cdots(\mathrm{mod}\ K) \tag{3.32}$$

where $\hat{\Delta}_i$, $i=1,2$, are as in (3.29).

(b) *For each small α there exists a small neighborhood N of $\Gamma_{12}\cup\Gamma_{21}\cup\{a_1,a_2\}$ and a K-homoclinic orbit γ to a_1 in N if and only if*

$$\begin{aligned}&\pi_{12}((0,y),\alpha)=\rho_2(\zeta_2^{(0)},\alpha),\\ &\pi_{21}(\rho_{22}(\zeta_2^{(j)},\alpha))=\rho_1(\zeta_1^{(j+1)},\alpha),\\ &\pi_{12}(\rho_{11}(\zeta_1^{(j+1)},\alpha),\alpha)=\rho_2(\zeta_2^{(j+1)},\alpha), \qquad j=0,1,2,\cdots(\mathrm{mod}\ K-1),\\ &\pi_{21}(\rho_{22}(\zeta_2^{(K-1)},\alpha),\alpha)=(x,0),\end{aligned} \tag{3.33}$$

have a solution with

$$\begin{aligned}&(0,x)\in\hat{\Sigma}_1^s(\alpha), \qquad (0,y)\in\hat{\Sigma}_1^u(\alpha),\\ &\zeta_i^{(j)}\in\hat{\Delta}_i(\alpha), \qquad i=1,2, \quad j=0,1,2,\cdots(\mathrm{mod}\ K-1)\end{aligned} \tag{3.34}$$

where $\hat{\Sigma}_1^s(\alpha)=\Sigma_1^s\cap N$ and $\hat{\Sigma}_1^u(\alpha)=\Sigma_1^u\cap N$.

Remark 3.4. (a) A statement similar to (b) of Lemma 3.3 holds true for homoclinic orbits to a_2. (b) An immediate consequence of this lemma is that to find periodic or homoclinic orbits near Γ we can first solve (3.31) and (3.33) without the constraints (3.32) and (3.34) associated with the small neighborhood N and then construct a neighborhood N independent of α such that (3.32) or (3.34) holds true.

We now introduce the changes of variables ρ_1 and ρ_2. Let $s_0>0$ be a constant. Let $\Delta_i=\Delta_i(s_0)$, $i=1,2$, be defined by

$$\Delta_i(s_0)=\{\zeta\in\mathbb{R}^{d-1}\mid \zeta=(s,x_0,y_1),\ s>s_0,\ (x_0,0)\in\Sigma_i^s,\ (0,y_1)\in\Sigma_i^u\}. \tag{3.35}$$

Let $(x,y)(t)=(x,y)(t;s,x_0,y_1,\alpha)$ be the solution of (3.11) satisfying (3.5). Define for every small α and $i=1,2$,

$$\rho_i(\cdot,\alpha):\Delta_i\to\Sigma_i^s \quad \text{with } (s,x_0,y_1)\to(x,y)(0). \tag{3.36}$$

We have the following lemma.

LEMMA 3.5. *There exists sufficiently large $S_0 > 0$ such that for all $s_0 > S_0$, ρ_i, defined by* (3.36), *defines a change of variables in σ_i^s. Moreover, the corresponding local map ρ_{ii}, defined by* (3.30), *is given by*

$$\rho_{ii}(\cdot, \alpha): \Delta_i \to \sigma_i^u \quad \text{with } (s, x_0, y_1) \to (x, y)(s) \tag{3.37}$$

where $i = 1, 2$.

Proof. It is obvious from Lemma 3.1 that ρ_i is differentiable. The exponential estimates in (3.7) and (3.8) imply that we can choose a sufficiently large $S_0 > 0$ such that $|y(0)| < \delta_{11}$ and $|x(s)| < \delta_{12}$ for all $s > s_0 \geqq S_0$ and $(s, x_0, y_1) \in \Delta_i$. Since both $(x, y)(0) \in \Sigma_i^s$ and $(x, y)(s) \in \Sigma_i^u$ are on the same orbit of (2.1), it follows from the definitions (3.17) and (3.18) for π_i, that $(x, y)(0) \in \sigma_i^s(\alpha)$ and $\pi_i(x(0), y(0), \alpha) = (x, y)(s) \in \sigma_i^u(\alpha)$. This, together with (3.30) implies (3.37).

To show that $\rho_i(\cdot, \alpha)$ is a diffeomorphism, recall the smooth scalar function s in the definitions (3.17) and (3.18) for the local map π_i. Note that

$$\bar{\rho}: \rho_i(\Delta_i, \alpha) \to \Delta_i$$

$$(x_0, y_0) \to (s, x_0, y(s; 0, x_0, y_0, \alpha)) \text{ with } s = s(x_0, y_0),$$

is actually the inverse for $\rho(\cdot, \alpha)$. Since $\bar{\rho}$ is also differentiable, $\rho(\cdot, \alpha)$ must be a local diffeomorphism. □

4. Uniqueness of homoclinic and periodic orbits. According to Lemma 3.3 and (b) of Remark 3.4, it suffices to consider (3.31) and (3.33) in the new variables in Δ_1 and Δ_2 introduced by (3.35) and (3.36). In this section, we will use these equations and prove the uniqueness of periodic and homoclinic orbits near $\Gamma_{12} \cup \Gamma_{21} \cup \{a_1, a_2\}$ (Proposition 4.1). To do so, we first rewrite the coordinates for Σ_i^s, Σ_i^u and Δ_i in some equivalent forms. Throughout this section, $i = 1$ or 2.

Since for all $(x, y) \in \Sigma_i^s$, $x^{(1)} = \delta_0$, there exists an obvious correspondence $(x, y) \leftrightarrow (\xi, y)$ with

$$\xi = (x^{(2)} - \bar{x}_{i0}^{(2)}, \cdots, x^{(m)} - \bar{x}_{i0}^{(m)}) \in \mathbb{R}^{m-1} \tag{4.1}$$

where $p_i = (\bar{x}_{i0}, 0)$ is as in (3.12). Note that $(\xi, y) \in \mathbb{R}^{d-1}$ and

$$\Sigma_i^s = \{(\xi, y) \in \mathbb{R}^{d-1} \,|\, |\xi| < \delta_{i1}, |y| < \delta_{1i}\}. \tag{4.2}$$

Similarly, let

$$\eta = (y^{(2)} - \bar{y}_{i1}^{(2)}, \cdots, y^{(n)} - \bar{y}_{i1}^{(n)}) \in \mathbb{R}^{n-1} \tag{4.3}$$

where $q_i = (0, \bar{y}_{i1})$ is as in (3.12). Then, (3.16) can be expressed as

$$\Sigma_i^u = \{(x, \eta) \in \mathbb{R}^{d-1} \,|\, |x| < \delta_{i2}, |\eta| < \delta_{i2}\} \tag{4.4}$$

(cf. Fig. 3.1).

Under the new variables (ξ, y) and (x, η) for Σ_i^s and Σ_i^u, respectively, $\pi_{12}: \Sigma_1^u \to \Sigma_2^s$ can be expressed as

$$\begin{aligned} \xi &= P_1(x, \eta, \alpha), \\ y &= Q_1(x, \eta, \alpha), \end{aligned} \qquad (x, \eta) \in \Sigma_1^u \tag{4.5}$$

where $\pi_{12} = (P_1, Q_1)$ is a diffeomorphism. Similarly, we have $\pi_{21} = (P_2, Q_2)$: $\Sigma_2^u \to \Sigma_1^s$.

In view of these changes of variables, the following correspondence is also a valid change of variables for both Δ_1 and Δ_2:

$$(s, x, y) \to (\tau, \xi, \eta) \tag{4.6}$$

where $(s, x, y) \in \Delta_i$ is as in (3.35), x and ξ, and y and η are related through (4.1) and (4.3), respectively, and

$$\tau = e^{-\mu(\alpha)s}. \tag{4.7}$$

Hence, (3.35) for Δ_i can be expressed as

$$\Delta_i = \{(\tau, \xi, \eta) \in \mathbb{R}^{d-1} \mid 0 < \tau < \tau_0, |\xi| < \delta_{i1}, |\eta| < \delta_{i2}\} \tag{4.8}$$

where

$$\tau_0 = e^{-\mu(\alpha)s_0}. \tag{4.9}$$

Note that Δ_i depends on the parameter α, which is suppressed for simplicity of notation.

In the new variables for Δ_i, let

$$X(\tau, \xi, \eta, \alpha) = x(s; s, x_0, y_1, \alpha), \tag{4.10}$$

$$Y(\tau, \xi, \mu, \alpha) = y(0; s, x_0, y_1, \alpha), \tag{4.11}$$

and

$$\psi(\xi, \eta, \alpha) = \varphi(x_0, y_1, \alpha) \tag{4.12}$$

where the solution $(x, y)(t) = (x, y)(t; s, x_0, y_1, \alpha)$ and the function φ are as in Theorem 3.1, and (τ, ξ, η) and (s, x_0, y_1) are related by (4.6) and (4.7). It is not difficult to see that from (3.7), (3.5), (4.7), (4.10)-(4.12) there exists a smooth function $R = R(\tau, \xi, \eta, \alpha)$ such that

$$\left| \frac{\partial}{\partial \tau} D^j X \right| = O(\tau^\nu), \tag{4.13}$$

$$Y = \tau\psi(\xi, \eta, \alpha) + R(\tau, \xi, \eta, \alpha), \tag{4.14}$$

$$\psi(\xi, \eta, \alpha) = \delta_0 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + O(|\xi| + |\eta|)\delta_0 \tag{4.15}$$

and

$$\left| \frac{\partial}{\partial \tau} D^j R \right| = O(\tau^\nu). \tag{4.16}$$

Here, the differentiation operator D^j involves derivatives only with respect to ξ, η and α, $j \leqq k-3$, and the constant ν satisfies

$$0 < \nu < \min\left\{ \frac{\lambda(\alpha)}{\mu(\alpha)} - 1, \frac{\bar{\nu}}{\mu(\alpha)} \right\}, \qquad |\alpha| \ll 1 \tag{4.17}$$

where λ is as in (3.6) and $\bar{\nu}$ is as in (3.8). To avoid confusion, we write X_i, Y_i, ψ_i, and R_i to denote the functions in (4.10), (4.11), (4.12), and (4.14), respectively, from Δ_i.

An important observation from (4.13)-(4.16) is that the functions X and Y can be C^1 extended to $\tau \leqq 0$. For simplicity, let X and Y denote such extensions in the extended region:

$$\tilde{\Delta}_i = \{(\tau, \xi, \eta) \in \mathbb{R}^{d-1} \mid |\tau| < \tau_0, |\xi| < \delta_{i1}, |\eta| < \delta_{i2}\}. \tag{4.18}$$

Obviously,

$$\Delta_i = \tilde{\Delta}_i \cap \{(\tau, \xi, \eta) \in \mathbb{R}^{d-1} \mid \tau > 0\}. \tag{4.19}$$

Note that both Δ_i and $\tilde{\Delta}_i$ depend on α. But, for the simplicity of notation, it is suppressed from the arguments.

Now, let us consider (3.31) and (3.33) for the periodic and homoclinic orbits near $\Gamma_{12} \cup \Gamma_{21}$. Let $K \geqq 1$, and $\zeta \in (\Delta_1 \times \Delta_2)^K$ with

$$\zeta = (\tau_1^0, \xi_1^0, \eta_1^0, \tau_2^0, \xi_2^0, \eta_2^0, \cdots, \tau_i^{K-1}, \xi_1^{K-1}, \eta_1^{K-1}, \tau_2^{K-1}, \xi_2^{K-1}, \eta_2^{K-1}). \tag{4.20}$$

Let

$$X_i^j = X(\tau_i^j, \xi_i^j, \eta_i^j, \alpha) \quad \text{and} \quad Y_i^j = Y(\tau_i^j, \xi_i^j, \eta_i^j, \alpha) \tag{4.21}$$

where

$$(\tau_i^j, \xi_i^j, \eta_i^j) \in \Delta_i, \qquad i = 1, 2, \quad j = 0, 1, 2, \cdots. \tag{4.22}$$

Then, it is not difficult to see that in the new variables τ, ξ, and η, equation (3.31) for periodic orbits is equivalent to

$$\Phi(\zeta, \alpha) = 0 \tag{4.23}$$

where

$$\Phi(\zeta, \alpha) = \begin{pmatrix} -\xi_2^j + P_1(X_1^j, \eta_1^j, \alpha) \\ -Y_2^j + Q_1(X_1^j, \eta_1^j, \alpha) \\ -\xi_1^{j+1} + P_2(X_2^j, \eta_2^j, \alpha) \\ -Y_1^{j+1} + Q_2(X_2^j, \eta_2^j, \alpha) \end{pmatrix}, j = 0, 1, 2, \cdots (\text{mod } K). \tag{4.24}$$

In particular, the estimates (4.13) and (4.16), (4.21), and (4.24) imply that (3.33) for the K-homoclinic orbits to a_1 is actually equivalent to (4.23) with $\zeta \in (\tilde{\Delta}_1 \times \tilde{\Delta}_2)^K$ satisfying

$$\tau_1^0 = 0, \quad (0, \xi_1^0, \eta_1^0) \in \tilde{\Delta}_1, \quad (\tau_2^0, \xi_2^0, \eta_2^0) \in \Delta_2 \tag{4.25}$$

and

$$(\tau_i^j, \xi_i^j, \eta_i^j) \in \Delta_i, \qquad i = 1, 2, \quad j = 1, 2, \cdots, K-1. \tag{4.26}$$

This shows that we only need to treat (4.23) uniformly in the extended domain $(\tilde{\Delta}_1 \times \tilde{\Delta}_2)^K$ for both the homoclinic and periodic orbits. Moreover, due to the following results (Theorem 4.2) on the uniqueness of periodic and homoclinic orbits in a small neighborhood N, we will see that we actually only have to consider (4.23) with $K = 1$. In this case, the correspondence between a solution $(\tau_1, \xi_1, \eta_1, \tau_2, \xi_2, \eta_2)$ of $\Phi = 0$ and an orbit are as follows:

(a) $\tau_1 = \tau_2 = 0$ corresponds to the heteroclinic loop from a_1 to a_2 and from a_2 to a_1,

(b) $\tau_1 = 0$ and $\tau_2 > 0$ corresponds to a homoclinic orbit from a_1 to a_1,

(c) $\tau_1 > 0$ and $\tau_2 = 0$ corresponds to a homoclinic orbit from a_2 to a_2,

(d) $\tau_1 > 0$ and $\tau_2 > 0$ corresponds to a periodic orbit.

In this section we will show that $\Phi = 0$ always has a unique solution for small α in the extended domain by Implicit Function Theorem arguments. However, it is obvious to see that the existence of solutions to $\Phi = 0$ does not always imply the

constraints $\tau_1 \geqq 0$ and $\tau_2 \geqq 0$ as required. Therefore, to get $\tau_1 \geqq 0$ and $\tau_2 \geqq 0$ for certain parameters, we need to derive some bifurcation equations from (4.23). This is to be done in the next section. We now have our results (Proposition 4.1 and Theorem 4.2) on the uniqueness of homoclinic and periodic orbits near $\Gamma_{12} \cup \Gamma_{21} \cup \{a_1, a_2\}$.

PROPOSITION 4.1. *Suppose* (2.1) *satisfies* (H1)-(H3) *and* (4.23) *satisfies the following conditions in the extended domain* $(\tilde{\Delta}_1 \times \tilde{\Delta}_2)^K$:

(4.27) *If* $\zeta^* = (\tau_1^*, \xi_1^*, \eta_1^*, \tau_2^*, \xi_2^*, \eta_2^*) \in \Delta_1 \times \Delta_2$ *is a solution to* (4.23), *with* $K = 1$, *then* $(\zeta^*, \cdots, \zeta^*) \in (\tilde{\Delta}_1 \times \tilde{\Delta}_2)^K$ *is also a solution to* (4.23) *with* $K > 1$.

(4.28) *For every small* α, *with* $K = 1$, (4.29) *has a unique solution* ζ^*.

(4.29) *For every small* α *and every* $K > 1$, *the solution of* (4.23) *in* $(\tilde{\Delta}_1 \times \tilde{\Delta}_2)^K$ *is unique.*

Then, for every small α *there exists at most one periodic or homoclinic orbit near* $\Gamma_{12} \cap \Gamma_{21} \cup \{a_1, a_2\}$, *but not both. Moreover, only simple* $(K = 1)$ *periodic or homoclinic orbits can exist.*

Proof. Suppose there exist two orbits γ_1 and γ_2, each of which is a periodic or homoclinic orbit. Then, there exist $K_1 \geqq 1$ and $K_2 \geqq 1$ such that (4.23) has a solution ζ_1^* for K_1 and ζ_2^* for K_2 with ζ_1^* satisfying either (4.22), or (4.25) and (4.26). Let $\zeta^* \in \tilde{\Delta}_1 \times \tilde{\Delta}_2$ be the solution to (4.23) with $K = 1$ guaranteed by the hypothesis (4.28). Then, hypotheses (4.27) and (4.29) imply

$$\zeta_i^* = (\zeta^*, \cdots, \zeta^*) \in (\tilde{\Delta}_1 \times \tilde{\Delta}_2)^{K_i}, \qquad i = 1, 2. \tag{4.30}$$

This implies that ζ^* must satisfy either (4.22), or (4.25) and (4.26). That is, there exists a simple periodic or homoclinic orbit near $\Gamma_{12} \cup \Gamma_{21} \cup \{a_1, a_2\}$. Furthermore, (4.30) forces K_1, $K_2 = 1$, and $\zeta_1^* = \zeta_2^* = \zeta^*$. This completes the proof. □

Finally, we have Theorem 4.2.

THEOREM 4.2. *Suppose that hypotheses* (H1)-(H4), (H7), *and* (H8) *are satisfied. Then there exists a neighborhood* N *of* $\Gamma_{12} \cup \Gamma_{21} \cup \{a_1, a_2\}$ *and a neighborhood* O *if* $\alpha = 0$ *such that for all small* α, *there exists at most one homoclinic or periodic orbit in* N, *but not both. Moreover, only simple* $(K = 1)$ *homoclinic or periodic orbits can exist.*

Proof. Note that (4.27)-(4.29) and Remark 3.4(b) of Lemma 3.3 imply the existence of such a neighborhood N. Hence, to prove the theorem it suffices to verify (4.27)-(4.29).

Condition (4.27) simply follows the definition (4.24) for Φ. To verify (4.28) and (4.29), we apply the Implicit Function Theorem to (4.23).

Let $K = 1$. Then $\zeta = (\tau_1^0, \xi_1^0, \eta_1^0, \tau_2^0, \xi_2^0, \eta_2^0) \in \mathbb{R}^{2(d-1)}$. The existence of Γ_{12} and Γ_{21} when $\alpha = 0$ implies that

$$\Phi(0, 0) = 0. \tag{4.31}$$

Since Φ is C^1, a simple calculation together with the eigenvalue conditions (H1) and (H2) implies

$$\frac{\partial \Phi}{\partial \zeta}(0, 0) = \begin{bmatrix} 0 & 0 & \partial P_{10}/\partial \eta & 0 & -I_{m-1} & 0 \\ 0 & 0 & \partial Q_{10}/\partial \eta & -\psi_{20} & 0 & 0 \\ 0 & -I_{m-1} & 0 & 0 & 0 & \partial P_{20}/\partial \eta \\ -\psi_{10} & 0 & 0 & 0 & 0 & \partial Q_{20}/\partial \eta \end{bmatrix}$$

where $\partial P_{i0}/\partial \eta = \partial P_i/\partial \eta$ $(0, 0, 0)$, $\partial Q_{i0}/\partial \eta = \partial Q_i/\partial \eta$ $(0, 0, 0)$, $\psi_{i0} = \psi_i$ $(0, 0, 0)$, and $i = 1, 2$. This matrix becomes diagonal through permutation, making the first two column

blocks into the last two column blocks. Then, it is easy to have

$$\det \frac{\partial \Phi}{\partial \zeta}(0,0) = (-1)^m \det\left[\frac{\partial Q_{10}}{\partial \eta}, \psi_{20}\right] \det\left[\frac{\partial Q_{20}}{\partial \eta}, \psi_{10}\right]. \tag{4.32}$$

We will show that (4.32) does not vanish.

Observe that the column vectors of the $(d-1)\times(d-2)$ matrix

$$M_1 = \begin{bmatrix} \partial P_{10}/\partial \eta & -I_{m-1} \\ \partial Q_{10}/\partial \eta & 0 \end{bmatrix} \tag{4.33}$$

span the subspace

$$\operatorname{span}\{T_p W_2^s, T_p W_1^u\} \cap T_p \Sigma_2^s \tag{4.34}$$

where $p = \Gamma_{12} \cap \Sigma_2^s$ is as in (3.12). This is because $T_p W_2^s = \mathbb{R}^m \times \{0\}$ and $T_p W_1^u = \operatorname{Im}(D_y \pi_{12}(0,0)) \times \operatorname{span}\{T_p \Gamma_{12}\}$. Since W_2^s and W_1^u are in general position (see (H5)) and Σ_2^s is transverse to Γ_{12}, we have

$$\operatorname{rank} M_1 = d-2. \tag{4.35}$$

Moreover, if δ_0 is as in Theorem 3.1, then, by (3.10) and (4.15) and the strong inclination property (H8), the vector $(0, \psi_{20}) = (0, \psi_2(0,0,0)) \in \mathbb{R}^{d-1}$ is complementary to the intersection of $\operatorname{span}\{T_p W_2^s, T_p W_1^u\}$ with $T_p \Sigma_2^s$. This implies

$$\operatorname{rank}\left[M_1, \begin{bmatrix} 0 \\ \psi_{20} \end{bmatrix}\right]_{(d-1)\times(d-1)} = d-1 \tag{4.36}$$

and thus,

$$\det\left[M_1, \begin{bmatrix} 0 \\ \psi_{20} \end{bmatrix}\right] = (-1)^{(n+1)(m-1)} \det\left[\frac{\partial Q_{10}}{\partial \eta}, \psi_{20}\right] \neq 0. \tag{4.37}$$

Similarly, we can show that

$$\det\left[\frac{\partial Q_{20}}{\partial \eta}, \psi_{10}\right] \neq 0. \tag{4.38}$$

We now conclude from (4.32) that

$$\det \frac{\partial \Phi}{\partial \zeta}(0,0) \neq 0. \tag{4.39}$$

It follows from the Implicit Function Theorem that there exist neighborhoods O of $\xi = 0 \in \tilde{\Delta}_1 \times \tilde{\Delta}_2$ and V of $\alpha = (0,0)$ in $\mathbb{R}^2$, and a C^1 function $\zeta^* = \zeta^*(\alpha)$, $\alpha \in V$, with $\zeta^*(\alpha) \in O$ for all $\alpha \in V$ and

$$\zeta^*(0) = 0 \tag{4.40}$$

such that ζ^* is the unique solution to (4.23). Note that by the Implicit Function Theorem, we also have that

$$\det \frac{\partial \Phi}{\partial \zeta}(\zeta, \alpha) \neq 0, \qquad \zeta \in O, \quad \alpha \in V. \tag{4.41}$$

This proves (4.28).

Next, we verify (4.29). Let $K > 1$ and consider (4.23) in $(\tilde{\Delta}_1 \times \tilde{\Delta}_2)^K$. The same reasoning as for (4.31) yields

$$\Phi(0,0) = 0. \tag{4.42}$$

Also, since Φ is C^1, simple calculation yields

$$\det\frac{\partial\Phi}{\partial\zeta}(0,0)=\begin{vmatrix} 0 & 0 & * & 0 & -I & 0 & & & & & & \\ 0 & 0 & * & * & 0 & 0 & & & & & & \\ & & & 0 & 0 & * & 0 & -I & 0 & & & \\ & & & 0 & 0 & * & * & 0 & 0 & & & \\ & & & & & & & & 0 & 0 & * & 0 & -I & 0 \\ & & & & & & & & 0 & 0 & * & * & 0 & 0 \\ 0 & -I & 0 & & & & & & & & & 0 & 0 & * \\ * & 0 & 0 & & & & & & & & & 0 & 0 & * \end{vmatrix} \tag{4.43}$$

where all nonspecified entries are zero. Similarly, by permutation we obtain a diagonal matrix such that every block of the form

$$\begin{bmatrix} * & 0 & -I \\ * & * & 0 \end{bmatrix}$$

in the diagonal is one of the following:

$$\begin{bmatrix} \partial P_{i0}/\partial\eta & 0 & -I_{m-1} \\ \partial Q_{i0}/\partial\eta & -\psi_{j0} & 0 \end{bmatrix}, \qquad i,j=1,2, \quad i\neq j.$$

By direct computation

$$\det\frac{\partial\Phi}{\partial\zeta}(0,0)=(-1)^m\left(\det\left[\frac{\partial Q_{10}}{\partial\eta},\psi_{20}\right]\cdot\det\left[\frac{\partial Q_{20}}{\partial\eta},\psi_{10}\right]\right)^K\neq 0 \tag{4.44}$$

because of (4.37) and (4.38). Moreover, it is not hard to see that (4.41) implies

$$\det\frac{\partial\Phi}{\partial\zeta}(\zeta,\alpha)=0, \qquad \zeta\in O^k, \quad \alpha\in V. \tag{4.45}$$

Hence, by the Implicit Function Theorem, equation (4.23) with $K>1$ has a unique solution $\zeta^*=\zeta^*(\alpha)$ for every $\alpha\in V$, ζ^* is C^1, and $\zeta^*(\alpha)\in O^k$ for all $\alpha\in V$. This proves (4.29). □

5. Bifurcation equations and proof of Theorem 2.1. From Theorem 4.2 it follows that a homoclinic or periodic orbit must satisfy

$$\Phi(\zeta,\alpha)=0, \quad \zeta\in\hat{\Delta}_1\times\hat{\Delta}_2, \quad \alpha\in\mathbb{R}^2 \tag{5.1}$$

with $\zeta=(\tau_1,\xi_1,\eta_1,\tau_2,\xi_2,\eta_2)$, $\tau_i\in\mathbb{R}$, $\xi_i\in\mathbb{R}^{m-1}$, $\eta_i\in\mathbb{R}^{n-1}$, and

$$\Phi(\zeta,\alpha)=\begin{pmatrix} -\xi_2+P_1(X_1,\eta_1,\alpha) \\ -Y_2+Q_1(X_1,\eta_1,\alpha) \\ -\xi_1+P_2(X_2,\eta_2,\alpha) \\ -Y_1+Q_2(X_2,\eta_2,\alpha) \end{pmatrix}. \tag{5.2}$$

Here, $X_i=X(\tau_i,\xi_i,\eta_i,\alpha)$, $Y_i=Y(\tau_i,\xi_i,\eta_i)\in\hat{\Delta}_i$, and (P_i,Q_i) are as in § 4. Recall, from the discussion given before Proposition 4.1, that the existence of homoclinic and periodic solutions is equivalent to the existence of solutions $\zeta\in\hat{\Delta}_1\times\hat{\Delta}_2$ of (5.1) satisfying

$$\tau_1\geqq 0, \quad \tau_2\geqq 0, \quad \tau_1+\tau_2\neq 0. \tag{5.3}$$

Also, a solution ζ of (5.1) with $\tau_1=\tau_2=0$ corresponds to the existence of a heteroclinic loop. Hence, the modified condition of (5.3):

$$\tau_1\geqq 0 \quad \text{and} \quad \tau_2\geqq 0 \tag{5.4}$$

for solutions ζ of (5.1) is equivalent to the existence of a homoclinic, periodic, or heteroclinic loop near $\Gamma_{12} \cup \Gamma_{21} \cup \{a_1, a_2\}$. From Theorem 4.2 it follows that any solution to (5.1) for $\alpha \in V$ must be $\zeta^* = \zeta^*(\alpha) = (\tau_1^*, \xi_1^*, \eta_1^*, \tau_2^*, \xi_2^*, \eta_2^*) \in O$, where O, V, and ζ^* are as in (4.40). Therefore, we need to verify that there exists a subset $\bar{\Lambda} \subset V$ such that

$$\zeta^* \in \Omega \quad \text{if and only if } \alpha \in \bar{\Lambda}, \tag{5.5}$$

where

$$\Omega = O \cap \{\zeta \in \hat{\Delta}_1 \times \hat{\Delta}_2 \mid \zeta = (\tau_1, \zeta_1, \eta_1, \tau_2, \xi_2, \eta_2),\ \tau_1 \geqq 0,\ \tau_2 \geqq 0\}. \tag{5.6}$$

This will be true if the mapping $h: \alpha \to (\tau_1^*(\alpha), \tau_2^*(\alpha)) \in \mathbb{R}^2$ is a diffeomorphism near $\alpha = 0$. Indeed, this is what we will prove in this section.

It is difficult to prove the nonsingularity of the mapping h by directly working with τ_1^*, τ_2^* and (5.1). Instead, we solve (5.1) in a way different from § 4 so that the correspondence between τ_1^*, τ_2^* and α can be easily determined through some bifurcation equations. To do so, we need the following notation. For $y = (y^{(1)}, \cdots, y^{(n)}) \in \mathbb{R}^n$, let

$$\hat{y} = (y^{(2)}, \cdots, y^{(n)}) \in \mathbb{R}^{n-1}. \tag{5.7}$$

For $z = (x, y) \in \mathbb{R}^{m-1} \times \mathbb{R}^n$, let

$$\hat{z} = (x, \hat{y}) \in \mathbb{R}^{m-1} \times \mathbb{R}^{n-1}. \tag{5.8}$$

Let

$$\Phi_i = \begin{pmatrix} -\xi_j + P_i(X_i, \eta_i, \alpha) \\ -Y_j + Q_i(X_i, \eta_i, \alpha) \end{pmatrix}, \qquad i, j = 1, 2, \quad i \neq j \tag{5.9}$$

where X_i and Y_i are as in (5.2). Then, for Φ defined by (5.2), we have

$$\Phi = \begin{pmatrix} \Phi_1 \\ \Phi_2 \end{pmatrix}. \tag{5.10}$$

We first solve the equation

$$\Phi_1(\zeta, \alpha) = 0 \tag{5.11}$$

where Φ_1 is as in (5.9). The existence of Γ_{12} and Γ_{21} at $\alpha = 0$ implies

$$\Phi_1(0, 0) = 0. \tag{5.12}$$

Also, by (H5), the existence of the heteroclinic orbit Γ_{α_1} for $\alpha = (\alpha_1, 0)$ implies that (5.11) always has solutions η_1 and ξ_2 for $\tau_1 = \tau_2 = 0$ and $\alpha = (\alpha_1, 0)$. That is,

$$Q_1(0, \eta_1, (\alpha_1, 0)) = 0 \quad \text{for some } |\eta_1| \ll 1. \tag{5.13}$$

Also, let $\hat{\Phi}_1$ be given by (5.8). Then, from (5.12), we have that

$$\hat{\Phi}_1(0, 0) = 0. \tag{5.14}$$

In addition, similarly to the computation for (4.32), it is easy to see that

$$\frac{\partial \Phi_1}{\partial(\eta_1, \xi_2)}(0, 0) = \begin{bmatrix} \partial P_{10}/\partial \eta & -I_{m-1} \\ \partial Q_{10}/\partial \eta & 0 \end{bmatrix} = M_1 \tag{5.15}$$

where M_1, $\partial P_{10}/\partial \eta$, and $\partial Q_{10}/\partial \eta$ are as in (4.33). Thus, (4.35) implies

$$\operatorname{rank} \frac{\partial \Phi_1}{\partial(\eta_1, \xi_2)}(0, 0) = d - 2. \tag{5.16}$$

Moreover, it is not difficult to see from (H8) of the strong inclination property and (5.15) that

$$\operatorname{rank} \frac{\partial \hat{\Phi}_1}{\partial(\eta_1, \xi_2)}(0,0) = d-2. \tag{5.17}$$

Hence, by (5.14), (5.17), and the Implicit Function Theorem, there exist C^1 functions $a = a(z, \alpha) \in \mathbb{R}^{n-1}$ and $b = b(z, \alpha) \in \mathbb{R}^{m-1}$, defined for $z \in \mathbb{R}^d$, $\alpha \in \mathbb{R}^2$ with $|z| \ll 1$, $|\alpha| \ll 1$, such that

$$\hat{\Phi}_1(\zeta, \alpha) = 0, \quad |\zeta| \ll 1, \quad |\alpha| \ll 1, \tag{5.18}$$

if and only if $\zeta = \zeta_*(\chi, \alpha)$, where

$$\zeta_* = (\tau_1, \xi_1, \eta_1, \tau_2, \xi_2, \eta_2) \quad \text{with } \eta_1 = a(\chi, \alpha) \text{ and } \xi_2 = b(\chi, \alpha), \tag{5.19}$$

and

$$\chi = (\tau_1, \xi_1, \tau_2, \eta_2).$$

Moreover, the Implicit Function Theorem implies that

$$a(0,0) = 0 \quad \text{and} \quad b(0,0) = 0. \tag{5.20}$$

Now, by substituting $\zeta = \zeta_*(\chi, \alpha)$ into the remaining equation, the full system of equation (5.11) is equivalent to the equation

$$\Phi_1^{(m)}(\zeta_*(\chi, \alpha), \alpha) = 0. \tag{5.21a}$$

Next, we will derive an equation equivalent to (5.21a). To do so, we first note some properties for the solution of (5.18). It is easy to see from (5.9) and (5.18) that when $\tau_1 = \tau_2 = 0$, the solutions $\eta_1 = a(\chi, \alpha)|_{\tau_1=\tau_2=0}$ and $\xi_2 = b(\chi, \alpha)|_{\tau_1=\tau_2=0}$ do not depend on ξ_1 and η_2. Thus

$$a_\alpha = a(\chi, \alpha)|_{\tau_1=\tau_2=0} \quad \text{and} \quad b_\alpha = b(\chi, \alpha)_{\tau_1=\tau_2=0} \tag{5.21b}$$

are functions of the variable α alone. This observation is very important. Obviously, we also have

$$a - a_\alpha = O(|\tau_1| + |\tau_2|) \quad \text{and} \quad b - b_\alpha = O(|\tau_1| + |\tau_2|). \tag{5.22}$$

Moreover, when $\tau_1 = \tau_2 = 0$, (5.18) implies

$$\begin{pmatrix} -b_\alpha + P_{1\alpha} \\ \hat{Q}_{1\alpha} \end{pmatrix} = \hat{\Phi}(\zeta_*(\chi, \alpha), \alpha)\Big|_{\tau_1=\tau_2=0} = 0 \tag{5.23}$$

where $P_{1\alpha} = P_1(0, a_\alpha, \alpha)$ and $Q_{1\alpha} = Q_1(0, a_\alpha, \alpha)$. In particular,

$$\hat{Q}_{1\alpha} = 0, \qquad |\alpha| \ll 1. \tag{5.24}$$

Let $\partial P_{1\alpha}/\partial \eta = \partial P_1/\partial \eta\ (0, a_\alpha, \alpha)$, $\partial Q_{1\alpha}/\partial \eta = \partial Q_1/\partial \eta\ (0, a_\alpha, \alpha)$ and

$$M_{1\alpha} = \begin{bmatrix} \partial P_{1\alpha}/\partial \eta & -I_{m-1} \\ \partial Q_{1\alpha}/\partial \eta & 0 \end{bmatrix}. \tag{5.25}$$

Then, by continuous dependence on α, by (5.15), and by (5.16), we have

$$\operatorname{rank} M_{1\alpha} = d-2. \tag{5.26}$$

In particular, by (5.17), we have

$$\operatorname{rank} \frac{\partial Q_{1\alpha}}{\partial \eta} = n-1. \tag{5.27}$$

Now, by (5.18), (5.24), and (5.25), a simple computation yields

$$\det\left[M_{1\alpha}, \Phi_1(\zeta_*, \alpha) - M_{1\alpha}\begin{pmatrix} a-a_\alpha \\ b \end{pmatrix}\right] = (-1)^{mn-1} \det\left[\frac{\partial Q_{1\alpha}}{\partial \eta}\right] \Phi_1^{(m)}(\zeta_*, \alpha). \tag{5.28}$$

Hence, (5.27) and (5.28) imply that the full system of equation (5.21b) is equivalent to the following equation:

$$\det\left[M_{1\alpha}, \Phi_1(\zeta_*, \alpha) - M_{1\alpha}\begin{pmatrix} a-a_\alpha \\ b \end{pmatrix}\right] = 0 \tag{5.29}$$

where $a = a(\chi, \alpha)$, $b = b(\chi, \alpha)$, $\zeta^* = \zeta_*(\chi, \alpha)$, and a_α are as in (5.19) and (5.21a).

To simplify (5.29) further, we expand (P_1, Q_1) at $(x, \eta) = (0, a_\alpha)$ and ψ_2 at $(\xi, \eta) = (b_\alpha, 0)$, respectively. Thus, by (5.22) we obtain

$$\begin{pmatrix} P_1(X_1, a, \alpha) \\ Q_1(X_1, a, \alpha) \end{pmatrix} = \begin{pmatrix} P_{1\alpha} \\ Q_{1\alpha} \end{pmatrix} + \frac{\partial(P_1, Q_1)}{\partial(x, \eta)}(0, a_\alpha, \alpha)\begin{pmatrix} X_1 \\ a-a_\alpha \end{pmatrix} + O(|\tau_1|^2 + |\tau_2|^2), \tag{5.30}$$

$$\psi_2(b, \eta_2) = \psi_{2\alpha} + O(|\eta_2| + |\tau_1| + |\tau_2|) \tag{5.31}$$

where $\psi_{2\alpha} = \psi_2(b_\alpha, 0)$. Also, from (5.31), (4.14), and (4.16) we have

$$Y_2(\tau_2, b, \eta_2, \alpha) = \psi_{2\alpha}\tau_2 + O(|\eta_2|\,|\tau_2| + |\tau_1|^{1+\nu} + |\tau_2|^{1+\nu}) \tag{5.32}$$

where $\nu > 0$ is as in (4.16). Now, substituting (5.30) with $X_1 = X(\tau_1, \xi_1, a, \alpha)$ and (5.32) into (5.29), from (4.15) we obtain

$$\begin{aligned} &-\det\left[M_{1\alpha}, \begin{pmatrix} 0 \\ \psi_{2\alpha} \end{pmatrix}\right]\tau_2 + \det\left[M_{1\alpha}, \begin{pmatrix} P_{1\alpha} \\ Q_{1\alpha} \end{pmatrix}\right] \\ &\quad + O(|\eta_2|\,|\tau_2| + |\tau_1|^{1+\nu} + |\tau_2|^{1+\nu}) = 0. \end{aligned} \tag{5.33}$$

Note that by the continuous dependence on α and (4.36),

$$\det\left[M_{1\alpha}, \begin{pmatrix} 0 \\ \psi_{2\alpha} \end{pmatrix}\right] \neq 0, \qquad |\alpha| \ll 1. \tag{5.34}$$

Hence, (5.33) can be further simplified to

$$\tau_2 = c_2(\alpha_1, \alpha_2) + O(|\eta_2|\,|\tau_2| + \tau_1|^{1+\nu} + |\tau_2|^{1+\nu}) \tag{5.35}$$

where

$$c_2(\alpha_1, \alpha_2) = \det\left[M_{1\alpha}, \begin{pmatrix} P_{1\alpha} \\ Q_{1\alpha} \end{pmatrix}\right] \Big/ \det\left[M_{1\alpha}, \begin{pmatrix} 0 \\ \psi_{2\alpha} \end{pmatrix}\right]. \tag{5.36}$$

PROPOSITION 5.1. *In addition to* (H1)-(H4), (H7), *and* (H8) *as in Theorem* 4.2, *suppose* (H5) *and* (H6) *are also satisfied. Then for sufficiently small* α

$$c_2(\alpha_1, 0) = 0 \tag{5.37}$$

and

$$\frac{\partial c_2}{\partial \alpha_2}(0, 0) \neq 0 \tag{5.38}$$

where c_2 *is given by* (5.36).

Proof. If $\tau_1 = \tau_2 = 0$, then

$$\Phi_1(\zeta, \alpha)|_{\tau_1=\tau_2=0} = \begin{pmatrix} -\xi_2 + P_1(0, \eta_2, \alpha) \\ Q_1(0, \eta_2, \alpha) \end{pmatrix}. \tag{5.39}$$

Moreover, (5.11) and (5.35) are equivalent. Thus, (5.13) and (5.35) imply (5.37). To prove (5.38), we first conclude from (5.22) that

$$\det\left[M_{1\alpha}, \begin{pmatrix} P_{1\alpha} \\ Q_{1\alpha} \end{pmatrix}\right] = (-1)^{mn-1} \det\left[\frac{\partial Q_{1\alpha}}{\partial \eta}\right] Q_{1\alpha}^{(1)}. \tag{5.40}$$

Since $(0, a_\alpha) \in W_1^u(\alpha)$, $\pi_{12}(0, a_\alpha, \alpha) = (P_{1\alpha}, Q_{1\alpha}) \in W_1^u(\alpha)$. Hence, (5.24) and (2.11) with $M_1^u(\alpha) = \{\pi_{12}(0, \eta, \alpha) \,|\, |\eta| \ll 1\} \subset \Sigma_2^s$ and $M_2^s(\alpha) = \{(x, 0) \,|\, |x| \ll 1\} \subset \Sigma_2^s$, imply that

$$0 \leqq d_1(\alpha_1, \alpha_2) \leqq \inf_{z \in M_2^s(\alpha)} |(P_{1\alpha}, Q_{1\alpha}) - z| \leqq |Q_{1\alpha}^{(1)}|. \tag{5.41}$$

Moreover, by (H5)

$$0 < d_1(0, \alpha_2) \quad \text{for } \alpha_2 \neq 0. \tag{5.42}$$

This and (5.41) imply $Q_{1\alpha}^{(1)}$ at $\alpha = (0, \alpha_2)$ with $\alpha_2 \neq 0$ being nonzero. Suppose, $Q_{1\alpha}^{(1)} > 0$ for $\alpha = (0, \alpha_2)$, $\alpha_2 > 0$. Then, we conclude from the transverse crossing hypothesis (H6) and (5.41) that

$$0 < \lim_{\alpha_2 \to 0^+} \frac{d_1(0, \alpha_2)}{\alpha_2} \leqq \lim_{\alpha_2 \to 0^+} \frac{Q_{1\alpha}^{(1)}}{\alpha_2}, \qquad \alpha = (0, \alpha_2). \tag{5.43}$$

Since (5.37), (5.40), and (5.27) also imply $Q_{1\alpha}^{(1)} = 0$ when $\alpha = (0, \alpha_2)$, it follows that

$$\lim_{\alpha_2 \to 0^+} \frac{Q_{1\alpha}^{(1)}}{\alpha_2} = \frac{\partial Q_{1\alpha}^{(1)}}{\partial \alpha_2}\bigg|_{\alpha=0}, \qquad \alpha = (0, \alpha_2). \tag{5.44}$$

Now, it is easy to see that (5.38) follows from the quotient rule of differentiation, (5.40), (5.43), and (5.44). □

Now, from Proposition 5.1 and (5.35) we have the following lemma.

LEMMA 5.2. *Suppose hypotheses* (H1)-(H8) *are satisfied. Then there exist a* C^1 *function* $c_2 = c_2(\alpha_1, \alpha_2)$ *satisfying* (5.37) *and a* C^1 *function* $r_2 = r_2(\tau_1, \xi_1, \tau_2, \eta_2, \alpha)$ *satisfying*

$$|r_2| = O(|\eta_2||\tau_2| + |\tau_1|^{1+\nu} + |\tau_2|^{1+\nu}) \tag{5.45}$$

with $\nu > 0$ *such that* (5.11) *has a solution* $\zeta = (\tau_1, \xi_1, \eta_1, \tau_2, \xi_2, \eta_2)$ *with* $|\zeta| \ll 1$ *and* $|\alpha| \ll 1$ *if and only if*

$$\tau_2 = c_2(\alpha_1, \alpha_2) + r_2(\tau_1, \xi_1, \tau_2, \eta_2, \alpha). \tag{5.46}$$

Note that by applying Lemma 5.2 to equation $\Phi_2(\zeta, \alpha) = 0$, there also exist C^1 functions $c_1 = c_1(\alpha_1, \alpha_2)$ and $r_1 = r_1(\tau_2, \xi_2, \tau_1, \eta_1, \alpha)$ such that equation $\Phi_2(\zeta, \alpha) = 0$ is equivalent to

$$\tau_1 = c_1(\alpha_1, \alpha_2) + r_1(\tau_2, \xi_2, \tau_1, \eta_1, \alpha). \tag{5.47}$$

In particular, by the proof of Theorem 4.2, if $\zeta^*(\alpha) = (\tau_1^*, \xi_1^*, \eta_1^*, \tau_2^*, \xi_2^*, \eta_2^*)(\alpha)$ solves (5.1), it must also satisfy (5.46) and (5.47) by Lemma 5.2. That is,

$$\begin{aligned} \tau_1^* &= c_1(\alpha_1, \alpha_2) + r_1(\tau_1^*, \xi_1^*, \tau_2^*, \eta_2^*, \alpha) \\ \tau_2^* &= c_2(\alpha_1, \alpha_2) + r_2(\tau_2^*, \xi_2^*, \tau_1^*, \eta_1^*, \alpha). \end{aligned} \tag{5.48}$$

These equations are considered as *bifurcation equations* for (5.1). Now, we can easily derive the following from (5.48).

Proof of Theorem 2.1. Without loss of generality, let $V \subset \mathbb{R}^2$ be the same as in Theorem 4.2. Since $|\zeta^*(\alpha)| = O(|\alpha|)$, (5.45) and (5.48) imply that

$$\tau_1^* = c_1(\alpha_1, \alpha_2) + O(|\alpha|^{1+\nu}), \qquad \tau_2^* = c_2(\alpha_1, \alpha_2) + O(|\alpha|^{1+\nu}). \tag{5.49}$$

In addition, (5.37) and (5.38) imply that the change of parameters

$$\varepsilon = c(\alpha) = (c_1(\alpha_1, \alpha_2), c_2(\alpha_1, \alpha_2)) \tag{5.50}$$

is nonsingular in V and satisfies (2.14) of Therorem 2.1. This and (5.49) imply that

$$\frac{\partial(\tau_1^*, \tau_2^*)}{\partial(\varepsilon_1, \varepsilon_2)}\bigg|_{\varepsilon=0} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{5.51}$$

That is the map $h: \varepsilon \to (\tau_1^*, \tau_2^*)$, which is a local diffeomorphism. Thus $\bar{\Lambda}$ in (5.5) and (5.6) is now given as $h^{-1}(\{\tau_1^* \geqq 0, \tau_2^* \geqq 0\})$. In particular, the boundary $h^{-1}(\{\tau_1^* = 0, \tau_2^* > 0\})$ of $\bar{\Lambda}$ corresponds to a unique homoclinic orbit to a_1 and is given as follows:

$$0 = \varepsilon_1 + O(|\varepsilon|^{1+\nu})$$

by (5.49). It is obvious that this boundary curve, corresponding to $\mathscr{C}_1$ in our theorem, and (2.15) and (2.16) follow immediately. The interior of $\bar{\Lambda}$ yields (2.17). This completes the proof. □

6. Application. Consider the FitzHugh-Nagumo equations

$$u_t = u_{xx} + f(u) - w, \qquad w_t = \varepsilon(u - \gamma w). \tag{6.1}$$

We refer the readers to [4], [7], [10], and [13] for more details on these equations. In (6.1) ε and γ are positive constants with $0 < \varepsilon \ll 1$. For $f(u)$, we take

$$f(u) = u(1-u)(u-a), \qquad 0 < a < \tfrac{1}{2}.$$

A traveling wave solution of (6.1) is a bounded, nonconstant solution of the form

$$(u(x, t), w(x, t)) = (U(z), W(z)), \qquad z = x + \theta t, \quad \theta = \text{constant}. \tag{6.2}$$

By substituting (6.2) into (6.1), (U, W) satisfies the following system of ordinary differential equations:

$$U' = V, \quad V' = \theta V - f(U) + W, \quad W' = \varepsilon(U - \gamma W)/\theta. \tag{6.3}$$

If γ is large enough, then there exist three rest points as shown in Fig. 6.1.

In what follows we let $\varphi = (\mathscr{E}_1, \mathscr{E}_2)$ be the point of intersection shown in Fig. 6.1, and $\mathscr{E} = (\mathscr{E}_1, 0, \mathscr{E}_2)$ the corresponding rest point of (6.3). We take $\mathscr{O} = (0, 0, 0)$ as another rest point.

By a pulse we mean a solution $\Gamma_{\mathscr{O}} = \Gamma_{\mathscr{O}}(z)$ of (6.3) that satisfies

$$\lim_{|z| \to +\infty} \Gamma_{\mathscr{O}}(z) = \mathscr{O}.$$

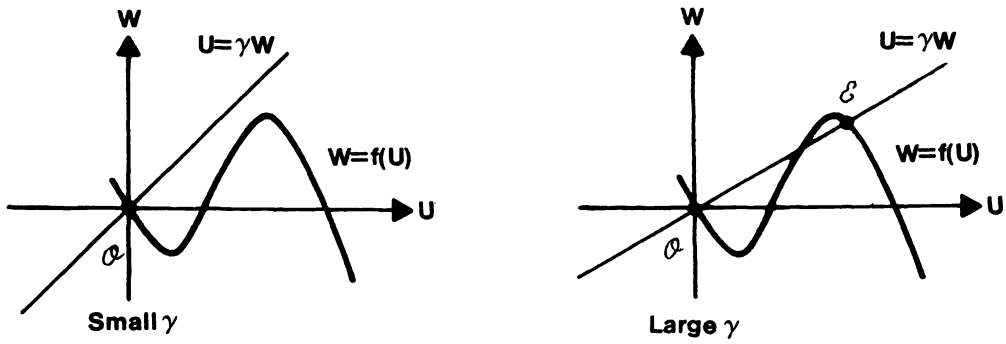


FIG. 6.1

By an $\mathscr{E}$-pulse $\Gamma_{\mathscr{E}}$ we mean a solution that satisfies

$$\lim_{|z|\to+\infty} \Gamma_{\mathscr{E}}(z) = \mathscr{E}.$$

By a front wave Γ_F we mean a solution that satisfies

$$\lim_{z\to+\infty} \Gamma_F(z) = \mathscr{E} \quad \text{and} \quad \lim_{z\to-\infty} \Gamma_F(z) = \mathscr{O}.$$

By a back wave Γ_B we mean a solution that satisfies

$$\lim_{z\to+\infty} \Gamma_B(z) = \mathscr{O} \quad \text{and} \quad \lim_{z\to-\infty} \Gamma_B(z) = \mathscr{E}.$$

Throughout this discussion we assume that $0 < \varepsilon \ll 1$. The relevant parameters are then γ and the wave speed θ. It has been shown (see [1], [10]) that there exist constants $0 < \gamma_1 < \gamma_2 < \gamma_3$ such that we have the following:

(1) If $\gamma_1 < \gamma$ and $0 < \varepsilon \ll 1$, then a front wave Γ_F exists for some θ, say $\theta_F(\gamma)$.

(2) If $\gamma_1 < \gamma < \gamma_3$ and $0 < \varepsilon \ll 1$, then a back wave Γ_B exists for some θ, say $\theta_B(\gamma)$.

(3) In the limit $\varepsilon \to 0$, the graphs of these differentiable functions $\theta_F(\gamma)$ and $\theta_B(\gamma)$ are approximately shown in Fig. 6.2.

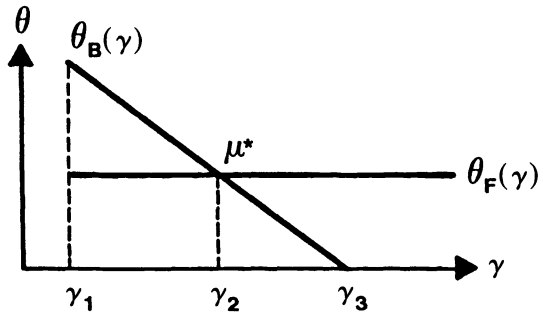


FIG. 6.2

These two curves cross precisely at $\gamma = \gamma_2$, in the limit $\varepsilon \to 0$. For ε small, but not zero, the curves $\theta_B(\gamma)$ and $\theta_F(\gamma)$ have the same qualitative features as shown in Fig. 6.2. For γ close to γ_1, $\theta_B(\gamma) > \theta_F(\gamma)$, while for γ close to γ_3, $\theta_B(\gamma) < \theta_F(\gamma)$. Hence there must exist $\mu^* = (\gamma^*(\varepsilon), \theta^*(\varepsilon))$ where the two curves cross.

It is not difficult to show that when $0 < \varepsilon \ll 1$, the linearizations of (5.3) at $\mathscr{O}$ and $\mathscr{E}$ have two negative eigenvalues and one positive eigenvalue. One of the negative eigenvalues is zero in the limit $\varepsilon \to 0$ while the other two stay uniformly away from zero as $\varepsilon \to 0$. Thus, (H1)-(H3) are satisfied only for the time-reverse system of (6.3). Hypothesis (H4) follows easily from (1)-(3) above. Hypothesis (H5) is always true for systems in $\mathbb{R}^3$. As mentioned earlier (H6)-(H8) are generic (see Fig. 6.3). We can verify these conditions do hold for (6.3) using the singular perturbation description of the wave (see [1] and [10]). Thus, Theorem 2.1 is applicable for the time-reverse system of (6.3).

Let $\theta_{\mathscr{O}}$ and $\theta_{\mathscr{E}}$ denote the pulse and ε-pulse curves in parameter space. Then, by Theorem 2.1 there are four possibilities for the location of the sector Λ that corresponds to the periodic orbits. However, Λ is determined by the following results taken from [1] and [10] together with the relative positions of θ_F, θ_B, $\theta_{\mathscr{O}}$, and $\theta_{\mathscr{E}}$ of Theorem 2.1:

(1) The front speed exceeds the pulse speed for the same γ;

(2) When both a front and back exist, a pulse exists only if the back speed exceeds the front speed. The corresponding statements also hold for ε-pulses (see Fig. 6.4).

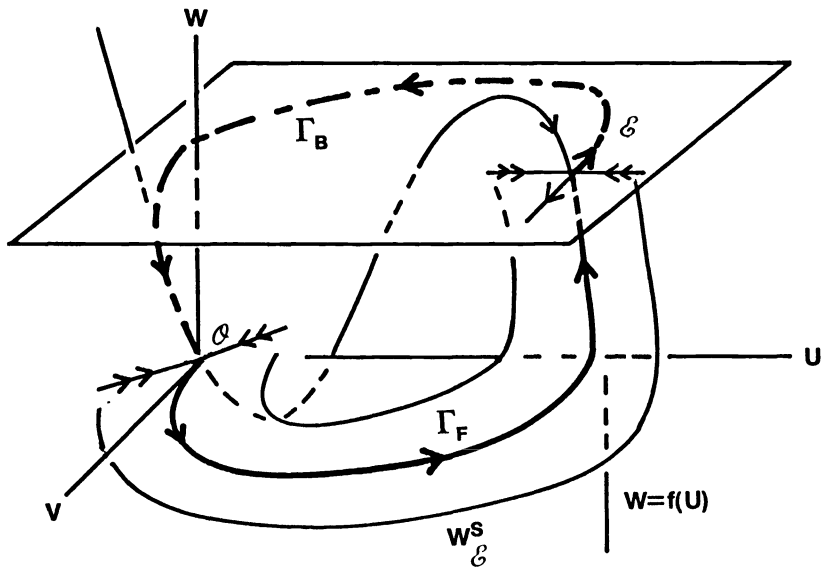


FIG. 6.3

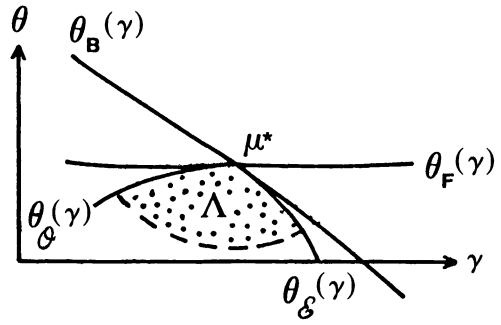


FIG. 6.4

The important consequence of this result is that for γ near γ_2 there exist infinitely many periodic waves traveling at speeds always exceeded by the pulse speed. This result is also contained in [1].

REFERENCES

[1] G. A. CARPENTER, *A geometric approach to singular perturbations problems with applications to nerve impulse equation*, J. Differential Equations, 23 (1977), pp. 335-367.

[2] S.-N. CHOW AND B. DENG, *Bifurcation of a unique and stable periodic orbit from a homoclinic orbit in some infinite dimensional systems*, Trans. Amer. Math. Soc., to appear.

[3] S.-N. CHOW, B. DENG, AND D. TERMAN, *The bifurcation of a homoclinic orbit from two heteroclinic orbits—a topological approach*, preprint.

[4] C. CONLEY, *Traveling wave solutions of nonlinear differential equations*, in Structural Stability, the Theory of Catastrophes, and Applications in the Sciences, A. Dold and B. Eckmann, eds., Lecture Notes in Mathematics, Springer-Verlag, Berlin, New York, 1975.

[5] B. DENG, *The Šil'nikov problem, exponential expansion, strong λ-lemma, C^1-linearization and homoclinic bifurcation*, J. Differential Equations, 79 (1989), pp. 189-231.

[6] J. K. HALE AND X.-B. LIN, *Heteroclinic orbits for retarded functional differential equations*, J. Differential Equations, 65 (1986), pp. 175-202.
[7] S. P. HASTINGS, *Single and multiple pulse nerves for the FitzHugh-Nagumo equations*, SIAM J. Appl. Math., 42 (1982), pp. 247-626.
[8] J. I. NEIMARK AND L. P. ŠIL'NIKOV, *A case of generation of periodic motions*, Sov. Math. Dokl., 6 (1965), pp. 1261-1264.
[9] K. J. PALMER, *Exponential dichotomies and transversal homoclinic points*, J. Differential Equations, 55 (1984), pp. 225-256.
[10] J. RINZEL AND D. TERMAN, *Propagation phenomena in a bistable reaction-diffusion system*, SIAM J. Appl. Math., 42 (1982), pp. 1111-1137.
[11] L. P. ŠIL'NIKOV, *On a Poincaré-Birkhoff problem*, Math. USSR-Sb., 3 (1967), pp. 353-371.
[12] ———, *On the generation of periodic motion from trajectories doubly asymptotic to an equilibrium state of saddle type*, Math. USSR-Sb., 77 (1968), pp. 427-438.
[13] D. TERMAN, *Threshold phenomena for a reaction-diffusion system*, J. Differential Equations, 47 (1983), pp. 406-443.
[14] P. FIFE, *Mathematical Aspect of Reaction Diffusion Systems*, Lecture Notes in Biomath. 25, Springer-Verlag, Berlin, New York, 1979.
[15] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, New York, 1983.
[16] E. YANAGIDA, *Branching of double pulse solutions from single pulse solutions in nerve axion equations*, J. Differential Equations, 66 (1986), pp. 243-262.
[17] S. N. CHOW, B. DENG, AND B. FIEDLER, *Homoclinic bifurcation at resonant eigenvalues*, preliminary report.
[18] L. P. ŠIL'NIKOV, *A case of the existence of a countable number of periodic motions*, Soviet Math. Dokl., 6 (1965), pp. 163-166.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 205-220, January 1990
© 1990 Society for Industrial and Applied Mathematics
011

LIPSCHITZ CONTINUOUS METRIC SELECTIONS IN $C_0(T)$*

WU LI†

Abstract. This paper gives an intrinsic characterization of those finite-dimensional subspaces G of $C_0(T)$ whose metric projections P_G have Lipschitz continuous selections. It is also proved that P_G has a Lipschitz continuous selection if and only if P_G is Lipschitz continuous.

Key words. Lipschitz continuity of metric projection, Lipschitz continuous selection of metric projection, Hausdorff strong unicity, extremal signature

AMS(MOS) subject classifications. 41A50, 41A65

1. Introduction. The problems concerning the existence of various continuous metric selections have received much attention in recent years. The existence of continuous metric selections is essential for finding stable algorithms to compute best approximations (cf. [24], [30]-[32]). Also, the existence of Lipschitz continuous metric selections and continuous metric selections will help us to determine the proximinality of certain tensor product subspaces in multivariate approximation (cf. [18] and "Sitting Duck Theorem" in [5], [19]). For finite-dimensional subspaces G of $C_0(T)$, the behavior of the metric projections P_G has been deeply investigated (cf. [8] and [23] for surveys). Nürnberger, Sommer, and Li (cf. [15], [17], [23], [8]) found several intrinsic characterizations of the existence of continuous selections for P_G. It was proved in [12] and [14], by Fischer and Li independently, that the almost lower semicontinuity of P_G is equivalent to the existence of a continuous selection for P_G. In [16], Li gives an intrinsic characterization of the lower semicontinuity of P_G. In [20], Lin generalizes a result by Deutsch [7] and establishes an intrinsic characterization of the existence of linear selections for P_G. Thus, as far as the various continuities of P_G are concerned, there is still one interesting question remaining: What are intrinsic characterizations of those G whose metric projections P_G have Lipschitz continuous selections (or whose P_G are Lipschitz continuous)?

In this paper, we will give the question above a complete answer. Before we go into detail, we introduce some notation.

Let T be a locally compact Hausdorff space and let $C_0(T)$ be the Banach space of real-valued continuous functions f on T that vanish at infinity, i.e., for any $\varepsilon > 0$, the set $\{t \in T: |f(t)| \geqq \varepsilon\}$ is compact. The norm of $f \in C_0(T)$ is defined as follows:

$$\|f\| = \sup\{|f(t)|: t \in T\}.$$

For $G \subset C_0(T)$, the metric projection from $C_0(T)$ to G is defined as

$$P_G(f) = \{g \in G: \|f - g\| = d(f, G)\},$$

where

$$d(f, G) = \inf\{\|f - p\|: p \in G\}.$$

Recall [13] that P_G is called Hausdorff strongly unique at $f \in C_0(T)$ if there is a constant $\lambda(f) > 0$ such that

$$\|f - g\| \geqq d(f, G) + \lambda(f) \cdot d(g, P_G(f)), \qquad g \in G.$$

* Received by the editors October 14, 1987; accepted for publication (in revised form) October 24, 1988.
† Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.

P_G is said to be uniform Hausdorff strongly unique if there is a constant $\lambda > 0$ such that

$$\|f-g\| \geqq d(f, G)+\lambda \cdot d(g, P_G(f)), \quad f \in C_0(T), \quad g \in G.$$

Remember that P_G is called Lipschitz continuous if there is a constant $\lambda > 0$ such that

$$H(P_G(f), P_G(h)) \leqq \lambda \cdot \|f-h\|, \qquad f, h \in C_0(T),$$

where

$$H(P_G(f), P_G(h)) = \max \{\sup \{d(g, P_G(f)): g \in P_G(h)\},$$
$$\sup \{d(p, P_G(h)): p \in P_G(f)\}\}.$$

We say that P_G has a Lipschitz continuous selection if there exists a Lipschitz continuous mapping Q from $C_0(T)$ to G such that $Q(f) \in P_G(f)$ for each $f \in C_0(T)$.

Now we can state our main results.

THEOREM 1.1. *Suppose that G is a finite-dimensional subspace of $C_0(T)$. Then the following are mutually equivalent:*

(i) *P_G is uniform Hausdorff strongly unique;*

(ii) *P_G is Lipschitz continuous;*

(iii) *P_G has a Lipschitz continuous selection;*

(iv) *$T \backslash Z(g)$ is compact for every $g \in G$, where $Z(g) = \{t \in T: g(t) = 0\}$.*

COROLLARY 1.2. *Suppose $c_0 = C_0(\mathbb{N})$ and G is a finite-dimensional subspace of c_0. Then the following are mutually equivalent:*

(1) *P_G is uniformly Hausdorff strongly unique;*

(2) *P_G is Lipschitz continuous;*

(3) *P_G has a Lipschitz continuous selection;*

(4) *There is $n \geqq 1$ such that $g(i) = 0$, $i \geqq n$, $g \in G$.*

Remark 1.3. Cline [6] knew that, for a compact T with finite points and a Haar subspace G of $C(T)$, P_G is Lipschitz continuous (cf. also [1]). Cline [6] also proved that for a compact T with infinite points and a Haar subspace G of $C(T)$, P_G is Lipschitz continuous only if $\dim G = 1$. The converse was proved by Berdyshev [2] (cf. also [28]). Respess and Cheney [28] extended Cline's result and proved that P_G has a Lipschitz continuous selection only if $\dim G = 1$, provided that G has Haar property at a neighborhood of a cluster point of T. We can easily derive these results by using Theorem 1.1. When T is compact, Berdyshev [2] also has some characterizations of P_G being Lipschitz continuous.

Remark 1.4. In general, there are one-dimensional subspaces of $C_0(T)$ for which the metric projections P_G have no continuous selections [3]. But for any finite-dimensional subspace G of c_0, P_G is lower semicontinuous (lsc) (cf. [4] or [16]). However, Corollary 1.2 implies that, in general, the lower semicontinuity of P_G is not equivalent to the Lipschitz continuity of P_G. Also, Corollary 1.2 implies that there is a one-dimensional subspace G of c_0 such that P_G is lsc but P_G has no Lipschitz continuous selection. It is interesting to note that metric projections in $L_1(T, \mu)$ have quite different features. Contrary to the phenomena mentioned above, for any one-dimensional subspace G of $L_1(T, \mu)$, P_G is lsc if and only if P_G is Lipschitz continuous and P_G has a continuous selection if and only if P_G has a Lipschitz continuous selection [9].

Remark 1.5. The idea of Hausdorff strong uniqueness is introduced in [13] to characterize the lower semicontinuity of P_G and is a natural generalization of the strong uniqueness of best approximations in $C_0(T)$ introduced by Newman and Shapiro [21]. However, it is interesting to note that the uniform Hausdorff strong uniqueness

may also be considered as a natural generalization of the $1\frac{1}{2}$-ball property introduced by Yost [33], since G has the $1\frac{1}{2}$-ball property if and only if [11]

$$\|f-g\| = d(f, G)+d(g, P_G(f)) \quad \text{for any } f\in C_0(T), \quad g\in G.$$

Some results in [33] have been generalized by Park [25] to the case that P_G is uniform Hausdorff strongly unique. Also, Park [25], [26] gives an example of a one-dimensional subspace G of $C[a, b]$ such that P_G is uniform Hausdorff strongly unique but G does not have the $1\frac{1}{2}$-ball property.

Remark 1.6. That (i) implies (ii) in Theorem 1.1 is the special case of Park's results [25] or [26]. By using the Steiner point, we can show that if P_G is Lipschitz continuous, then P_G has a Lipschitz continuous selection (cf. [10], [27]). Thus (ii) implies (iii).

Since a subset A of $\mathbb{N}$ is compact if and only if A is a finite subset of $\mathbb{N}$, Corollary 1.2 follows immediately from Theorem 1.1. By Remark 1.6, we need to show only that (iii) implies (ii), which in turn implies (i). In § 2 we give some properties of extremal signatures of G that will play an important role in this paper. In § 3 we show that (iii) implies (iv). In § 4 we study the structure of G that satisfies (iv). In § 5 we prove that (iv) implies (i).

2. Extremal signatures. A signature σ on T is a mapping from T to $\{-1, 0, 1\}$ such that $\{t\in T: \sigma(t)\neq 0\} =: \operatorname{supp}\sigma$ is a nonempty finite set. Extremal signatures were introduced by Rivlin and Shapiro [29] to characterize elements in $P_G(f)$. For an equivalent definition of extremal signatures see [17], [29].

DEFINITION 2.1. An extremal signature σ of G is a signature on T such that if $g\in G$ satisfies $\sigma(t)g(t)\geqq 0$, for $t\in\operatorname{supp}\sigma$, then $\operatorname{supp}\sigma\subset Z(g)$. A primitive extremal signature σ of G is an extremal signature of G such that

$$\dim G|_{\operatorname{supp}\sigma} = \dim G|_{\operatorname{supp}\sigma\setminus\{t\}} = \operatorname{card}(\operatorname{supp}\sigma)-1 \quad \text{for } t\in\operatorname{supp}\sigma.$$

Now we list some properties of extremal signatures of G that will be used in this paper.

LEMMA 2.2 (Rivlin and Shapiro [29]). *Suppose $f\in C_0(T)\setminus G$ and $g\in G$. Then $g\in P_G(f)$ if and only if there is a primitive extremal signature σ of G such that*

$$f(t)-g(t) = \sigma(t)\|f-g\| \quad \text{for } t\in\operatorname{supp}\sigma.$$

LEMMA 2.3. *Suppose that σ is an extremal signature of G. Then $P_{G|_{\operatorname{supp}\sigma}}(\sigma|_{\operatorname{supp}\sigma}) = \{0\}$ and there is a constant $\lambda(\sigma)>0$ such that*

$$\max\{\sigma(t)g(t): t\in\operatorname{supp}\sigma\}\geqq\lambda(\sigma)\cdot\max\{|g(t)|: t\in\operatorname{supp}\sigma\} \quad \text{for } g\in G.$$

LEMMA 2.4. *Suppose $B\subset T$. Then $\operatorname{card}(B)>\dim G|_B$ if and only if there is an extremal signature σ with $\operatorname{supp}\sigma\subset B$, where $\operatorname{card}(B)$ denotes the cardinal number of the set B.*

LEMMA 2.5. *If σ_1 is an extremal signature of G and σ_2 is an extremal signature of $G(\operatorname{supp}\sigma_1) := \{g\in G: \operatorname{supp}\sigma_1\subset Z(g)\}$, then*

$$\sigma(t) = \begin{cases}\sigma_1(t), & t\in\operatorname{supp}\sigma_1,\\ \sigma_2(t), & t\in T\setminus\operatorname{supp}\sigma_1,\end{cases}$$

is an extremal signature of G.

Lemmas 2.3–2.5 are results in § 2 of [17]. By Lemmas 2.4 and 2.5 we can establish the following auxiliary lemma, which will be used in the proof that (iii) implies (iv) in Theorem 1.1.

LEMMA 2.6. *Suppose* $B \subset T$ *and* $\operatorname{card}(B) > \dim G|_B$. *Then there is an extremal signature* σ *of* G *such that* $\operatorname{supp}\sigma \subset B$ *and*

$$\dim G(\operatorname{supp}\sigma)|_B = \operatorname{card}(B\backslash Z(G(\operatorname{supp}\sigma))),$$

where $G(\operatorname{supp}\sigma) = \{g \in G: \operatorname{supp}\sigma \subset Z(g)\}$.

Proof. We prove this lemma by induction on $\dim G$. Assume that Lemma 2.6 is true if $\dim G \leqq s$, where $s \geqq 0$. Now suppose $\dim G = s+1$. If $\dim G|_B = \operatorname{card}(B\backslash Z(G))$, then there is $t_0 \in B \cap Z(G)$, since $\dim G|_B < \operatorname{card}(B)$. Obviously,

$$\sigma(t) = \begin{cases} 1, & t = t_0, \\ 0, & t \in T\backslash\{t_0\} \end{cases}$$

is an extremal signature of G and

$$\begin{aligned} \dim G(\operatorname{supp}\sigma)|_B &= \dim G(t_0)|_B = \dim G|_B \\ &= \operatorname{card}(B\backslash Z(G)) = \operatorname{card}(B\backslash Z(G(\operatorname{supp}\sigma))), \end{aligned}$$

since $G(\operatorname{supp}\sigma) = G$. Thus, without loss of generality, we may assume

$$\dim G|_B \neq \operatorname{card}(B\backslash Z(G)). \tag{2.1}$$

Since $\dim G|_B = \dim G|_{B\backslash Z(G)} \leqq \operatorname{card}(B\backslash Z(G))$, (2.1) implies

$$\dim G|_{B\backslash Z(G)} < \operatorname{card}(B\backslash Z(B)). \tag{2.2}$$

By Lemma 2.4, there is an extremal signature σ_1 of G with $\operatorname{supp}\sigma_1 \subset B\backslash Z(G)$. Set $G^* = G(\operatorname{supp}\sigma_1)$. Then $\dim G^* \leqq \dim G - 1 \leqq s$. Since $\dim G^*|_B \leqq \dim G|_B < \operatorname{card}(B)$, by the inductive hypothesis, there is an extremal signature σ_2 of G^* such that

$$\operatorname{supp}\sigma_2 \subset B, \qquad \dim G^*(\operatorname{supp}\sigma_2)|_B = \operatorname{card}(B\backslash Z(G^*(\operatorname{supp}\sigma_2))).$$

Define

$$\sigma(t) = \begin{cases} \sigma_1(t), & t \in \operatorname{supp}\sigma_1, \\ \sigma_2(t), & t \in T\backslash\operatorname{supp}\sigma_1. \end{cases}$$

By Lemma 2.5, σ is an extremal signature of G. Since

$$\operatorname{supp}\sigma = \operatorname{supp}\sigma_1 \cup \operatorname{supp}\sigma_2,$$

we obtain that

$$\operatorname{supp}\sigma \subset B, \tag{2.3}$$

$$\begin{aligned} \dim G(\operatorname{supp}\sigma)|_B &= \dim G^*(\operatorname{supp}\sigma_2)|_B \\ &= \operatorname{card}(B\backslash Z(G^*(\operatorname{supp}\sigma_2))) = \operatorname{card}(B\backslash Z(G(\operatorname{supp}\sigma))). \end{aligned} \tag{2.4}$$

Formulae (2.3) and (2.4) show that σ is the required extremal signature of G.

3. Proof of (iii) implying (iv). Suppose that P_G has a Lipschitz continuous selection Q. If (iv) in Theorem 1.1 fails to be true, then there is $g^* \in G$ such that $\operatorname{supp}(g^*) = T\backslash Z(g^*)$ is not compact. Set

$$T_k = \{t \in T: 0 < |g^*(t)| < 1/k\}, \qquad k \geqq 1.$$

If for some k, T_k is a finite set, then

$$\operatorname{supp}(g^*) = \{t \in T: |g^*(t)| \geqq \varepsilon\}, \tag{3.1}$$

where

$$\varepsilon = \min\{|g^*(t)|: t \in T_k\} > 0. \tag{3.2}$$

Formulae (3.1) and (3.2) imply that supp (g^*) is compact. This contradicts our assumption that supp (g^*) is not compact. Thus, for each $k \geqq 1$, T_k is an infinite set. Since $T_k \supset T_{k+1}$, $k \geqq 1$, we have

$$0 \leqq \dim G|_{T_{k+1}} \leqq \dim G|_{T_k} \leqq \dim G < +\infty, \qquad k \geqq 1,$$

i.e., $\{\dim G|_{T_k}\}_{k=1}^{\infty}$ is a bounded decreasing sequence. So, there is $r \geqq 0$ such that

$$\lim_{k\to\infty} \dim G|_{T_k} = r. \tag{3.3}$$

Since $\dim G|_{T_k}$ are integers, (3.3) implies that there is $N > 0$ such that

$$\dim G|_{T_k} = \dim G|_{T_N} = r, \qquad k \geqq N. \tag{3.4}$$

It follows from (3.4) that

$$\begin{aligned} \dim G(T_k) &= \dim G - \dim G|_{T_k} \\ &= \dim G - \dim G|_{T_N} = \dim G(T_N), \qquad k \geqq N, \end{aligned} \tag{3.5}$$

where $G(T_k) = \{g \in G\colon T_k \subset Z(g)\}$. By (3.5) and $G(T_k) \supset G(T_N)$ for $k \geqq N$, we obtain

$$G(T_k) = G(T_N), \qquad k \geqq N. \tag{3.6}$$

For each $k \geqq N$, since T_k is an infinite set, by Lemma 2.6 there is an extremal signature σ_k of G such that

$$\operatorname{supp} \sigma_k \subset T_k, \tag{3.7}$$

$$\operatorname{card}(T_k \backslash Z(G_k)) \leqq \dim G_k \leqq \dim G < +\infty, \tag{3.8}$$

where

$$G_k = \{g \in G\colon \operatorname{supp} \sigma_k \subset Z(g)\}. \tag{3.9}$$

It follows from (3.7) and (3.9) that

$$G_k \supset G(T_N), \qquad k \geqq N. \tag{3.10}$$

Set

$$\varepsilon_k = \min\{|g^*(t)|\colon t \in T_k \backslash Z(G_k)\}. \tag{3.11}$$

By (3.8), $\varepsilon_k > 0$. By (3.11) and the definition of T_k, we obtain

$$T_n \cap (T_k \backslash Z(G_k)) = \varnothing, \quad n > 1/\varepsilon_k, \quad n \geqq k,$$

i.e.,

$$T_n \subset Z(G_k), \quad n > 1/\varepsilon_k, \quad n \geqq k. \tag{3.12}$$

Formula (3.12) implies

$$G(T_n) \supset G_k, \quad n > 1/\varepsilon_k, \quad n \geqq k. \tag{3.13}$$

It follows from (3.6), (3.10), and (3.13) that

$$G_k = G(T_N), \qquad k \geqq N. \tag{3.14}$$

Since supp σ_k is a finite set and T_k is open, there are open sets W_k and V_k such that

$$\operatorname{supp} \sigma_k \subset V_k \subset \bar{V}_k \subset W_k \subset \bar{W}_k \subset T_k, \qquad \bar{W}_k \text{ is compact,}$$

where $\bar{A}$ denotes the closure of set A. Now, by using Tietze's extension theorem, we can construct f_k, $h_k \in C_0(T)$ such that

$$
\begin{aligned}
&f_k(t) = \sigma_k(t), \qquad t \in \operatorname{supp} \sigma_k,\\
&\|f_k\| = 1,\\
&f_k(t) = 0, \qquad t \in T \setminus V_k;\\
&h_k(t) = g^*(t), \qquad t \in \bar{V}_k,\\
&\|h_k\| = \max\{|g^*(t)|: t \in \bar{V}_k\} < 1/k,\\
&h_k(t) = 0, \qquad t \in T \setminus W_k.
\end{aligned}
$$

Let $0 < \alpha < \min\{\frac{1}{2}, \frac{1}{2}\|g^*\|\}$. Then

$$
\begin{aligned}
1 &= \max\{|f_k(t)|: t \in \operatorname{supp} \sigma_k\}\\
&= \max\{|f_k(t) + \alpha h_k(t) - \alpha g^*(t)|: t \in \operatorname{supp} \sigma_k\}\\
&\leqq \|f_k + \alpha h_k - \alpha g^*\|\\
&\leqq \max\{\max\{|f_k(t)|: t \in \bar{V}_k\}, \max\{\alpha |h_k(t) - g^*(t)|: t \in \bar{W}_k\}\}\\
&\leqq \max\{1, \alpha(\|h_k\| + \|g^*\|)\}\\
&\leqq \max\{1, \alpha(1/k + \|g^*\|)\} \leqq 1, \qquad k \geqq N,
\end{aligned}
$$

i.e.,

$$
\|f_k + \alpha h_k - \alpha g^*\| = 1. \tag{3.15}
$$

By (3.15) and the construction of f_k, h_k, we get

$$
\begin{aligned}
&f_k(t) = \sigma_k(t) = \sigma_k(t)\|f_k\| \quad \text{for } t \in \operatorname{supp} \sigma_k,\\
&f_k(t) + \alpha h_k(t) - \alpha g^*(t) = \sigma_k(t) = \sigma_k(t)\|f_k - \alpha h_k - \alpha g^*\| \quad \text{for } t \in \operatorname{supp} \sigma_k.
\end{aligned}
$$

If $g \in P_G(f_k)$, then

$$
|\sigma_k(t) - g(t)| = |f_k(t) - g(t)| \leqq \|f_k\| = 1 \quad \text{for } t \in \operatorname{supp} \sigma_k.
$$

By Lemma 2.3, $g|_{\operatorname{supp}\sigma_k} \in P_{G|_{\operatorname{supp}\sigma k}}(\sigma|_{\operatorname{supp}\sigma_k})$ and $g(t) = 0$ for $t \in \operatorname{supp} \sigma_k$, i.e.,

$$
P_G(f_k) \subset G_k, \qquad k \geqq N. \tag{3.16}
$$

Similarly, we can obtain

$$
P_G(f_k + \alpha h_k) - \alpha g^* = P_G(f_k + \alpha h_k - \alpha g^*) \subset G_k, \qquad k \geqq N. \tag{3.17}
$$

Thus, by (3.14), for the Lipschitz continuous selection Q of P_G, we have

$$
\begin{aligned}
&Q(f_k) \in G(T_N), \qquad k \geqq N,\\
&Q(f_k + \alpha h_k) - \alpha g^* \in G(T_N), \qquad k \geqq N.
\end{aligned}
$$

So, for $k \geqq N$,

$$
\begin{aligned}
\alpha \cdot \max\{|g^*(t)|: t \in T_N\} &= \max\{|Q(f_k, t) - Q(f_k + \alpha h_k, t)|: t \in T_N\}\\
&\leqq \|Q(f_k) - Q(f_k + \alpha h_k)\| \leqq \lambda \cdot \|f_k - f_k - \alpha h_k\| \qquad (3.18)\\
&= \lambda \cdot \alpha \cdot \|h_k\| < \lambda \cdot \alpha / k,
\end{aligned}
$$

where $\lambda > 0$ is the constant in the definition of Lipschitz continuity of Q. It follows from (3.18) that

$$0 < \max\{|g^*(t)|: t \in T_N\} < \lambda/k, \qquad k \geqq N,$$

which is impossible. The contradiction shows that (iii) implies (iv) in Theorem 1.1.

4. An equivalent form of (iv). We first assume that G satisfies (iv) in Theorem 1.1. Set

$$\operatorname{supp}(G) = \{t \in T: g(t) \neq 0 \text{ for some } g \in G\}.$$

We define a relation on $\operatorname{supp}(G)$ as follows:

$$x \sim y \quad \text{iff } \dim G|_{\{x,y\}} = 1,$$

where $x, y \in \operatorname{supp}(G)$. Then it is easy to check that "$\sim$" is an equivalence relation on $\operatorname{supp}(G)$. Let

$$\mathscr{X} = \{[x]: x \in \operatorname{supp}(G)\},$$

where $[x]$ denotes the equivalence class of x.

LEMMA 4.1. $\operatorname{supp}(G)$ *is compact and open.*

Proof. Suppose $G = \operatorname{span}\{g_i\}_1^r$. Then

$$\operatorname{supp}(G) = \bigcup_{i=1}^{r} \operatorname{supp}(g_i),$$

where $\operatorname{supp}(g_i) = \{t \in T: g_i(t) \neq 0\}$. By the continuity of g_i, $\operatorname{supp}(g_i)$ is open; by (iv) and $\operatorname{supp}(g_i) = T \backslash Z(g_i)$, $\operatorname{supp}(g_i)$ is compact. Thus, $\operatorname{supp}(G)$ is open and compact.

LEMMA 4.2. $[x]$ *is open and compact for each* $x \in \operatorname{supp}(G)$.

Proof. Let $\{g_i\}_1^s \subset G$ such that

$$\operatorname{span}\{g_i\}_1^s = \{g \in G: g(x) = 0\} =: G(x).$$

Since $y \in Z(G(x)) \cap \operatorname{supp}(G)$ if and only if $y \in \operatorname{supp}(G)$ and $\dim G|_{\{x,y\}} = 1$, where $Z(G(x)) = T \backslash \operatorname{supp}(G(x))$, we obtain

$$[x] = \operatorname{supp}(G) \cap Z(G(x)). \tag{4.1}$$

Since $Z(G(x)) = \{t \in T: g(t) = 0 \text{ for all } g \in G(x)\}$ is closed, $[x]$ is a closed subset of the compact set $\operatorname{supp}(G)$. So, $[x]$ is compact. On the other hand, by (iv), $Z(g_i) = T \backslash \operatorname{supp}(g_i)$ is open for each $1 \leqq i \leqq s$, so $Z(G(x)) = Z(\operatorname{span}\{g_i\}_1^s) = \bigcap_{i=1}^s Z(g_i)$ is open. By Lemma 4.1 and (4.1), we have that $[x]$ is open. Thus, $[x]$ is open and compact.

LEMMA 4.3. $\mathscr{X}$ *is a finite set.*

Proof. Assume the contrary, i.e., $\mathscr{X}$ is an infinite set. Then there are $\{x_k\}_1^\infty \subset \operatorname{supp}(G)$ such that

$$[x_i] \neq [x_j], \quad i \neq j, \quad i, j \geqq 1. \tag{4.2}$$

By Lemma 4.1, $\operatorname{supp}(G)$ is compact. Let $x^* \in \operatorname{supp}(G)$ be a cluster point of $\{x_k\}_1^\infty$. Since, by Lemma 4.2, $[x^*]$ is open, we obtain that $[x^*] \cap \{x_k\}_1^\infty$ is an infinite set. Let $i \neq j$, $x_i, x_j \in [x^*]$. Then

$$[x_i] = [x^*] = [x_j],$$

which contradicts (4.2). So, $\mathscr{X}$ is finite.

THEOREM 4.4. *Suppose that G is a finite-dimensional subspace of $C_0(T)$. Then the following are equivalent*:

(a) *$T\backslash Z(g)$ is compact for every $g \in G$;*

(b) *There are open and compact subsets $\{A_i\}_1^n$ such that*

$$\bigcup_{i=1}^{n} A_i = \operatorname{supp}(G), \tag{4.3}$$

and for $1 \leqq i, j \leqq n$,

$$\dim G|_{A_i \cup A_j} = 1 \quad \text{only if } i = j. \tag{4.4}$$

Proof. Part (a) implies (b). By Lemma 4.3, there are $\{x_i\}_1^n \subset \operatorname{supp}(G)$ such that

$$[x_i] \neq [x_j], \quad i \neq j, \quad 1 \leqq i, \quad j \leqq n, \tag{4.5}$$

$$\bigcup_{i=1}^{n} [x_i] = \bigcup_{x \in \operatorname{supp}(G)} [x] = \operatorname{supp}(G). \tag{4.6}$$

But $\dim G|_{[x_i] \cup [x_j]} = 1$ implies $[x_i] = [x_j]$. So, by (4.5), we get

$$\dim G|_{[x_i] \cup [x_j]} = 1 \quad \text{iff } i = j. \tag{4.7}$$

From (4.1) we obtain

$$G(x) = \{g \in G: g|_{[x]} \equiv 0\} =: G([x]), \qquad x \in \operatorname{supp}(G). \tag{4.8}$$

Hence, by (4.8) we get

$$1 = \dim G|_{\{x\}} = \dim G - \dim G(x) = \dim G - \dim G([x]) = \dim G|_{[x]}, \quad x \in \operatorname{supp}(G). \tag{4.9}$$

It follows that

$$\dim G|_{[x_i]} = 1, \qquad 1 \leqq i \leqq n. \tag{4.10}$$

Now, (4.6), (4.7), (4.10), and Lemma 4.2 ensure that (b) holds for $A_i = [x_i]$, $1 \leqq i \leqq n$.

Part (b) implies (a). By (4.4), we know that for any $g \in G$, $Z(g) \cap A_i \neq \varnothing$ implies $A_i \subset Z(g)$. Thus, for any $g \in G$, there is a subset J of $\{i\}_1^n$ such that $T\backslash Z(g) = \bigcup_{i \in J} A_i$ is compact.

5. Proof of (iv) implying (i). To prove that (iv) implies (i), we need several lemmas.

LEMMA 5.1. *For any $B \subset T$, there is a constant $\lambda(B) > 0$ such that*

$$d(g, G(B)) \leqq \lambda(B) \|g\|_B, \qquad g \in G,$$

where $\|g\|_B = \sup\{|g(t)|: t \in B\}$.

Proof. Let $\{g_i\}_1^r \subset G$ be such that $G|_B = \operatorname{span}\{g_i|_B\}_1^r$ and $\{g_i|_B\}_1^r$ is linearly independent. Let

$$G^* = \operatorname{span}\{g_i\}_1^r.$$

Then it is not difficult to check that $\|\cdot\|_B$ and $d(\cdot, G(B))$ are two norms on G^*. Since any two norms on a finite-dimensional space are equivalent, there is a constant $\lambda(B) > 0$ such that

$$d(g^*, G(B)) \leqq \lambda(B) \|g^*\|_B, \qquad g^* \in G^*. \tag{5.1}$$

Now let $g \in G$. Since $G^*|_B = G|_B$, there is $g^* \in G^*$ such that

$$g|_B = g^*|_B, \tag{5.2}$$

i.e.,

$$g - g^* \in G(B). \tag{5.3}$$

By (5.1)-(5.3) we get

$$d(g, G(B)) = d(g^*, G(B)) \leqq \lambda(B)\|g^*\|_B = \lambda(B) \cdot \|g\|_B, \qquad g \in G.$$

LEMMA 5.2. *Suppose $f \in C_0(T)$, $T = B_1 \cup B_2$, $B_1 \subset Z(G)$, and $G^* = G|_{B_2}$. If there is a positive constant α_1 such that*

$$\|f-g\| \geqq d(f|_{B_2}, G^*) + \alpha_1 \cdot d(g|_{B_2}, P_{G^*}(f|_{B_2})), \qquad g \in G, \tag{5.4}$$

then for every $g \in G$,

$$\|f-g\| \geqq d(f, G) + \left(\frac{\alpha_1}{2}\right) \cdot d(g, P_G(f)). \tag{5.5}$$

Proof. Suppose $g \in G$ and

$$\|f-g\| = d(f, G) + \mu. \tag{5.6}$$

If $\mu \leqq 0$, $g \in P_G(f)$ and (5.5) is trivial. So we may assume $\mu > 0$. Let $g^* \in G$ such that $g^*|_{B_2} \in P_{G^*}(f|_{B_2})$ and

$$d(g|_{B_2}, P_{G^*}(f|_{B_2})) = \|g - g^*\|_{B_2}. \tag{5.7}$$

For $0 < \lambda \leqq 1$, we claim

$$\|f - g^* + \lambda(g^* - g)\|_{B_2} \leqq d(f|_{B_2}, G^*) + \lambda \cdot (d(f, G) - d(f|_{B_2}, G^*) + \mu). \tag{5.8}$$

In fact, if (5.8) fails to be true, then there is $t_\lambda \in B_2$ such that

$$|f(t_\lambda) - g^*(t_\lambda) + \lambda(g^*(t_\lambda) - g(t_\lambda))| > d(f|_{B_2}, G^*) + \lambda \cdot \mu^*, \tag{5.9}$$

where

$$\mu^* = d(f, G) - d(f|_{B_2}, G^*) + \mu > 0.$$

Since $|f(t_\lambda) - g^*(t_\lambda)| \leqq d(f|_{B_2}, G^*) < d(f|_{B_2}, G^*) + \lambda \cdot \mu^*$, by (5.9), we obtain

$$|f(t_\lambda) - g^*(t_\lambda) + \lambda(g^*(t_\lambda) - g(t_\lambda))| = \varepsilon_\lambda(f(t_\lambda) - g^*(t_\lambda)) + \lambda|g^*(t_\lambda) - g(t_\lambda)|, \tag{5.10}$$

where $\varepsilon_\lambda = \operatorname{sign}(g^*(t_\lambda) - g(t_\lambda))$. Thus, by (5.9) and (5.10), we have

$$\begin{aligned}
\|f-g\| &\geqq \varepsilon_\lambda(f(t_\lambda) - g(t_\lambda)) \\
&= \varepsilon_\lambda(f(t_\lambda) - g^*(t_\lambda)) + |g^*(t_\lambda) - g(t_\lambda)| \\
&> \varepsilon_\lambda(f(t_\lambda) - g^*(t_\lambda)) + \frac{1}{\lambda} d(f|_{B_2}, G^*) + \lambda\mu^* - \varepsilon_\lambda(f(t_\lambda) - g^*(t_\lambda)) \\
&= \left(1 - \frac{1}{\lambda}\right) \varepsilon_\lambda \cdot (f(t_\lambda) - g^*(t_\lambda)) + \frac{1}{\lambda} d(f|_{B_2}, G^*) + \mu^* \\
&\geqq \left(1 - \frac{1}{\lambda}\right) \|f - g^*\|_{B_2} + \frac{1}{\lambda} d(f|_{B_2}, G^*) + \mu^* \\
&= \left(1 - \frac{1}{\lambda}\right) d(f|_{B_2}, G^*) + \frac{1}{\lambda} d(f|_{B_2}, G^*) + \mu^* \\
&= d(f, G) + \mu,
\end{aligned}$$

which contradicts (5.6). The contradiction shows that our claim (5.8) is true.

Set $\lambda^* = 1 - \mu/\mu^*$. Then, by (5.8), we get

$$\begin{aligned}\|f - g^* + \lambda^*(g^* - g)\| &\leqq d(f|_{B_2}, G^*) + \lambda^* \cdot \mu^* \\ &= d(f|_{B_2}, G^*) + \mu^* - \mu = d(f, G),\end{aligned}$$

i.e.,

$$g^* + \lambda^*(g - g^*) \in P_G(f).$$

So,

$$d(g, P_G(f)) \leqq \|g - g^* - \lambda^*(g - g^*)\| = (1 - \lambda^*)\|g - g^*\|. \tag{5.11}$$

We discuss two cases.

Case (a). $d(g|_{B_2}, P_{G^*}(f|_{B_2})) \geqq 2(\mu^* - \mu)/\alpha_1$.

By the hypothesis, we have

$$\begin{aligned}\|f - g\| &\geqq d(f|_{B_2}, G^*) + \alpha_1 \cdot d(g|_{B_2}, P_{G^*}(f|_{B_2})) \\ &\geqq d(f|_{B_2}, G^*) + \mu^* - \mu + \frac{\alpha_1}{2} \cdot d(g|_{B_2}, P_{G^*}(f|_{B_2})) \\ &= d(f, G) + \frac{\alpha_1}{2} \cdot d(g|_{B_2}, P_{G^*}(f|_{B_2})).\end{aligned} \tag{5.12}$$

Since $T = B_1 \cup B_2$, $B_1 \subset Z(G)$, and $G^* = G|_{B_2}$, it is easy to verify that

$$P_{G^*}(f|_{B_2}) \subset P_G(f)|_{B_2}, \tag{5.13}$$

$$d(g|_{B_2}, P_G(f)|_{B_2}) = d(g, P_G(f)). \tag{5.14}$$

It follows from (5.12)-(5.14) that

$$\|f - g\| \geqq d(f, G) + \frac{\alpha_1}{2} \cdot d(g, P_G(f)).$$

Case (b). $d(g|_{B_2}, P_{G^*}(f|_{B_2})) < 2(\mu^* - \mu)/\alpha_1$.

By (5.7) and $T \setminus B_2 \subset B_1 \subset Z(G)$, we have

$$d(g|_{B_2}, P_{G^*}(f|_{B_2})) = \|g - g^*\|_{B_2} = \|g - g^*\|. \tag{5.15}$$

It follows from (5.15), (5.11), and the hypothesis that

$$\begin{aligned}d(g, P_G(f)) &\leqq (1 - \lambda^*)\|g - g^*\| \\ &= d(g|_{B_2}, P_{G^*}(f|_{B_2})) \cdot \mu/\mu^* \\ &\leqq 2 \cdot (\mu^* - \mu) \cdot \mu/(\mu^* \cdot \alpha_1) \\ &\leqq 2\mu/\alpha_1.\end{aligned} \tag{5.16}$$

By (5.6) and (5.16), we obtain

$$\|f - g\| \geqq d(f, G) + \frac{\alpha_1}{2} \cdot d(g, P_G(f)).$$

Thus, we have proved that (5.5) holds for any $g \in G$.

LEMMA 5.3. *Suppose* $f \in C_0(T)$, $T = B_1 \cup B_2$, *and* $G^* = G(B_1)|_{B_2}$. *If there are positive constants* α_1, α_2 *such that*

$$\|f - g\| \geqq d(f, G) + \alpha_1 \cdot d(g|_{B_1}, P_G(f)|_{B_1}), \qquad g \in G, \tag{5.17}$$

$$\|f-g-g^*\| \geqq d(f-g|_{B_2}, G^*)+\alpha_2 \cdot d(g^*|_{B_2}, P_{G^*}(f-g|_{B_2})), \quad g \in P_G(f), \quad g^* \in G^*, \tag{5.18}$$

then

$$\|f-g\| \geqq d(f, G)+\alpha_3 \cdot d(g, P_G(f)), \qquad g \in G, \tag{5.19}$$

where $\alpha_3=\min\{\alpha_2/4, \alpha_1 \cdot \alpha_2/[2 \cdot \lambda(B_1) \cdot (\alpha_2+2)]\}$ *and* $\lambda(B_1)$ *is any constant that satisfies*

$$d(g, G(B_1)) \leqq \lambda(B_1) \cdot \|g\|_{B_1}, \qquad g \in G.$$

Proof. Suppose $g \in G$. Let $g_1 \in P_G(f)$ such that

$$\|g-g_1\|_{B_1}=d(g|_{B_1}, P_G(f)|B_1). \tag{5.20}$$

By Lemma 5.1, there is a positive constant $\lambda(B_1)$ such that

$$d(p, G(B_1)) \leqq \lambda(B_1)\|p\|_{B_1}, \qquad p \in G. \tag{5.21}$$

By (5.20) and (5.21), there is $g_2 \in G(B_1)$ such that

$$\|g-g_1-g_2\|=d(g-g_1, G(B)) \leqq \lambda(B_1)\|g-g_1\|_{B_1}=\lambda(B_1)d(g|_{B_1}, P_G(f)|_{B_1}). \tag{5.22}$$

Now we discuss two cases.

Case (a). $d(g|_{B_1}, P_G(f)|_{B_1}) \leqq \alpha_2 \cdot d(g, P_G(f))/[2 \cdot \lambda(B_1) \cdot (\alpha_2+2)]$.

By (5.18) and Lemma 5.2, we get

$$\|f-g_1-g_2\| \geqq d(f-g_1, G(B_1))+\frac{\alpha_2}{2} \cdot d(g_2, P_{G(B_1)}(f-g_1)). \tag{5.23}$$

Since $g_1 \in P_G(f)$, we have

$$d(f-g_1, G(B))=d(f, G), \tag{5.24}$$

$$P_{G(B_1)}(f-g_1) \subset P_G(f-g_1)=P_G(f)-g_1. \tag{5.25}$$

It follows from (5.23)-(5.25) that

$$\begin{aligned}\|f-g_1-g_2\| &\geqq d(f, G)+\frac{\alpha_2}{2} d(g_2, P_G(f)-g_1) \\ &= d(f, G)+\frac{\alpha_2}{2} d(g_1+g_2, P_G(f)).\end{aligned} \tag{5.26}$$

Since $d(g, P_G(f)) \geqq d(g_1+g_2, P_G(f))-\|g-g_1-g_2\|$, (5.22) and (5.26) imply

$$\begin{aligned}\|f-g\| &\geqq \|f-g_1-g_2\|-\|g-g_1-g_2\| \\ &\geqq d(f, G)+\frac{\alpha_2}{2} d(g, P_G(f))-\left(1+\frac{\alpha_2}{2}\right) \cdot \|g-g_1-g_2\| \\ &\geqq d(f, G)+\frac{\alpha_2}{2} d(g, P_G(f))-\lambda(B_1) \cdot \left(1+\frac{\alpha_2}{2}\right) \cdot d(g|_{B_1}, P_G(f)|_{B_1}).\end{aligned} \tag{5.27}$$

Now it follows from (5.27) and the hypothesis that

$$\|f-g\| \geqq d(f, G)+\frac{\alpha_2}{4} d(g, P_G(f)). \tag{5.28}$$

Case (b). $d(g|_{B_1}, P_G(f)|_{B_1}) > \alpha_2 \cdot d(g, P_G(f))/[2 \cdot \lambda(B_1) \cdot (\alpha_2+2)]$.

By (5.17) and the hypothesis, we obtain

$$\begin{aligned} \|f-g\| &\geqq d(f, G)+\alpha_1 \cdot d(g|_{B_1}, P_G(f)|_{B_1}) \\ &\geqq d(f, G)+\alpha_1\alpha_2 \cdot d(g, P_G(f))/[2 \cdot \lambda(B_1) \cdot (\alpha_2+2)]. \end{aligned} \tag{5.29}$$

Thus, for $\alpha_3 = \min\{\alpha_2/4, \alpha_1 \cdot \alpha_2/[2 \cdot \lambda(B_1) \cdot (\alpha_2+2)]\}$, by (5.28) and (5.29), (5.19) holds.

From now on, we will always assume that G satisfies (iv) in Theorem 1.1. By Theorem 4.4, there are open and compact subsets $\{A_i\}_1^n$ such that

$$\bigcup_{i=1}^{n} A_i = \operatorname{supp}(G), \tag{5.30}$$

and for $1 \leqq i, j \leqq n$,

$$\dim G|_{A_i \cup A_j} = 1 \quad \text{iff } i = j. \tag{5.31}$$

For convenience, denote

$$A(I) = \bigcup_{i \in I} A_i, \qquad I \subset \{1, 2, \cdots, n\}. \tag{5.32}$$

To give a lower-bounded estimation of Hausdorff strongly unique constants, we need some structural constants of G. By (5.31), there are $g_i \in G$ such that

$$G|_{A_i} = \operatorname{span}\{g_i|_{A_i}\}, \qquad 1 \leqq i \leqq n. \tag{5.33}$$

Since $g_i(t) \neq 0$ for $t \in A_i$ and A_i is compact, we obtain

$$\alpha_i =: \|g_i\|_{A_i}/\min\{|g_i(t)|: t \in A_i\} < \infty, \qquad 1 \leqq i \leqq n. \tag{5.34}$$

Define

$$1 \leqq \alpha = \max\{\alpha_i: 1 \leqq i \leqq n\} < \infty. \tag{5.35}$$

For any $g \in G$, if $g|_{A_i} \not\equiv 0$, then $g|_{A_i} = \lambda_i g_i|_{A_i}$ with $\lambda_i \neq 0$. So,

$$\|g\|_{A_i}/\min\{|g(t)|: t \in A_i\} = \alpha_i \leqq \alpha, \qquad g \in G \text{ with } \|g\|_{A_i} \neq 0. \tag{5.36}$$

Set

$$\mathcal{T} = \{I \subset \{i\}_1^n: \dim G|_{A(I)} = \dim G|_{A(I\setminus\{j\})} = \operatorname{card}(I) - 1 \text{ for } j \in I\}.$$

For $I \in \mathcal{T}$, it is easy to check that $\|\cdot\|_{A(I)}$ and $\|\cdot\|_{A(I\setminus\{j\})}$, $j \in I$, are norms on $G|_{A(I)}$. Since any two norms on $G|_{A(I)}$ are equivalent, there is a constant $\beta(I) > 0$ such that

$$\|g\|_{A(I\setminus\{j\})} \geqq \beta(I) \cdot \|g\|_{A(I)}, \quad g \in G, \quad j \in I \in \mathcal{T}. \tag{5.37}$$

Define

$$1 \geqq \beta = \min\{\beta(I): I \in \mathcal{T}\} > 0, \tag{5.38}$$

since $\mathcal{T}$ is a finite set. By Lemma 5.1, for $\varnothing \neq I \subset \{i\}_1^n$, there are constants $\gamma(I)$ such that

$$d(g, G(A(I))) \leqq \gamma(I)\|g\|_{A(I)}, \quad g \in G, \quad I \subset \{i\}_1^n. \tag{5.39}$$

Define

$$1 \leqq \gamma = \max\{\gamma(I): I \subset \{i\}_1^n\} < \infty. \tag{5.40}$$

The constants α, β, γ are essential in estimating the lower bound of Hausdorff strongly unique constants of P_G.

LEMMA 5.4. *For any primitive extremal signature σ of G and any real number $\lambda \geqq 0$, we have*

$$\|\lambda\sigma - g\|_{\operatorname{supp}\sigma} \geqq \|\lambda\sigma\|_{\operatorname{supp}\sigma} + (\beta/\alpha)\|g\|_{\operatorname{supp}\sigma}, \qquad g \in G. \tag{5.41}$$

Proof. Since $\beta/\alpha \leqq 1$, (5.41) is trivial if $\lambda = 0$. So, we may assume

$$\lambda > 0.$$

Set $G^* = G|_{\operatorname{supp}\sigma}$, $\sigma^* = \sigma|_{\operatorname{supp}\sigma}$, $\operatorname{supp}\sigma = \{t_i\}_0^m$. By Definition 2.1 and Lemma 2.3, we obtain that

$$P_{G^*}(\lambda\sigma^*) = \lambda P_{G^*}(\sigma^*) = \{0\}, \tag{5.42}$$

$$\dim G^* = \dim G|_{\{t_i\}_0^m} = \dim G|_{\{t_i\}_0^m \setminus \{t_j\}} = m, \qquad 0 \leqq j \leqq m. \tag{5.43}$$

It is trivial that

$$|\lambda\sigma^*(t)| = \|\lambda\sigma^*\|, \qquad t \in \operatorname{supp}\sigma = \{t_i\}_0^m. \tag{5.44}$$

So, it follows from (5.42)-(5.44) and Theorem 1.3 of [22] that

$$\|\lambda\sigma^* - g^*\| \geqq \|\lambda\sigma^*\| + \eta \cdot \|g^*\|, \qquad g^* \in G^*, \tag{5.45}$$

where

$$\eta = \min\{1/\|p_i^*\| : 0 \leqq i \leqq m\} \tag{5.46}$$

and p_i^* are the unique functions in G^* such that

$$p_i^*(t_j) = \operatorname{sign}(\lambda\sigma^*(t_j)) = \sigma^*(t_j), \quad 0 \leqq i, \quad j \leqq m, \quad j \neq i. \tag{5.47}$$

Choose $p_i \in G$ such that

$$p_i|_{\operatorname{supp}\sigma} = p_i^*, \qquad 0 \leqq i \leqq m. \tag{5.48}$$

Let $\{k_i : 0 \leqq i \leqq m\} \subset \{1, \cdots, n\}$ such that

$$t_i \in A_{k_i}, \qquad 0 \leqq i \leqq m. \tag{5.49}$$

By (5.31), (5.43), and (5.49), we obtain

$$\dim G|_{\cup_{i=0}^m A_{k_i}} = \dim G|_{\cup_{i=0}^m A_{k_i} \setminus A_{k_j}} = m, \qquad 0 \leqq j \leqq m. \tag{5.50}$$

By (5.47), (5.48), and (5.36), we can derive that

$$\|p_i\|_{A_{k_j}} \leqq \alpha \cdot |p_i(t_j)| = \alpha \cdot |p_i^*(t_j)| = \alpha, \quad 0 \leqq i, \quad j \leqq m, \quad i \neq j. \tag{5.51}$$

Now we discuss two cases.

Case (a). $m = 1$.

By (5.50) and (5.31), $A_{k_0} = A_{k_1}$. So, it follows from (5.51) that

$$\|p_i^*\| = \|p_i\|_{\{t_0, t_1\}} \leqq \|p_i\|_{A_{k_i}} \leqq \alpha, \qquad i = 0, 1.$$

So,

$$\eta \geqq \min\{1/\|p_i^*\| : i = 0, 1\} \geqq 1/\alpha. \tag{5.52}$$

Case (b). $m > 1$.

By (5.50) and $m > 1$, we obtain that for $I = \{k_i : 0 \leqq i \leqq m\}$,

$$\dim G|_{A(I)} = \dim G|_{A(I \setminus \{j\})} = \operatorname{card}(I) - 1 = m, \qquad j \in I;$$

i.e., $I \in \mathcal{T}$. It follows from (5.51) that

$$\|p_i\|_{A(I \setminus \{k_i\})} \leqq \alpha, \qquad 0 \leqq i \leqq m. \tag{5.53}$$

By (5.53), (5.37), and (5.38), we get

$$\|p_i\|_{A(I)} \leqq \|p_i\|_{A(I\setminus\{k_i\})}/\beta \leqq \alpha/\beta, \qquad 0 \leqq i \leqq m.$$

Thus,

$$\begin{aligned} \eta &\geqq \min\{1/\|p_i^*\|: 0 \leqq i \leqq m\} = \min\{1/\|p_i\|_{\operatorname{supp}\sigma}: 0 \leqq i \leqq m\} \\ &\geqq \min\{1/\|p_i\|_{A(I)}: 0 \leqq i \leqq m\} \geqq \beta/\alpha. \end{aligned} \tag{5.54}$$

Since $\beta \leqq 1$, (5.52) and (5.54) imply

$$\eta \geqq \beta/\alpha. \tag{5.55}$$

So, by (5.45) and (5.55), we obtain that for any $g \in G$,

$$\begin{aligned} \|\lambda\sigma - g\|_{\operatorname{supp}\sigma} &= \|\lambda\sigma^* - g|_{\operatorname{supp}\sigma}\| \\ &\geqq \|\lambda\sigma^*\| + \eta \cdot \|g|_{\operatorname{supp}\sigma}\| \geqq \|\lambda\sigma^*\| + (\beta/\alpha)\|g|_{\operatorname{supp}\sigma}\| \\ &= \|\lambda\sigma\|_{\operatorname{supp}\sigma} + (\beta/\alpha) \cdot \|g\|_{\operatorname{supp}\sigma}. \end{aligned}$$

This shows that (5.41) holds.

5.1. Proof that (iv) implies (i) in Theorem 1.1. We prove that (iv) implies (i) by induction on dim G. The conclusion is trivial if dim $G = 0$. Assume that for any subspace M with dim $M \leqq s$, if M satisfies (iv), then P_M is uniform Hausdorff strongly unique. Now suppose that G satisfies (iv) and dim $G = s+1$.

Let $T^* = \operatorname{supp}(G)$, $G^* = G|_{T^*}$. Set

$$G_I = G^*(A(I))|_{T^*\setminus A(I)}, \qquad \varnothing \neq I \subset \{i\}_1^n.$$

Since $G_1 \subsetneqq G^*$ and T^* is compact, we obtain that G_I satisfies (iv) and dim $G_I \leqq$ dim $G^* - 1 \leqq$ dim $G - 1 = s$. By the inductive hypothesis, P_{G_I} is uniform Hausdorff strongly unique, i.e., there is a positive constant $\eta(I) > 0$ such that

$$\|h - p\| \geqq d(h, G_I) + \eta(I) \cdot d(p, P_{G_I}(h)), \quad h \in C(T^*\setminus A(I)), \quad p \in G_I.$$

Since $\{i\}_1^n$ has only finitely many different subsets, we have

$$\eta =: \min\{\eta(I): \varnothing \neq I \subset \{i\}_1^n\} > 0.$$

Obviously, for any nonempty set $I \subset \{i\}_1^n$,

$$\|h - p\| \geqq d(h, G_I) + \eta \cdot d(p, P_{G_I}(h)), \quad p \in G_I, \quad h \in C(T^*\setminus A(I)). \tag{5.56}$$

Now we claim

$$\|f - g\| \geqq d(f, G) + \lambda \cdot d(g, P_G(f)), \quad g \in G, \quad f \in C_0(T), \tag{5.57}$$

where

$$\lambda = \min\{\eta/8, \beta \cdot \eta/[\alpha^2 \cdot \gamma \cdot 4(\eta+2)]\}.$$

First assume $f^* = f|_{T^*} \in C(T^*)\setminus G^*$. By Lemma 2.2, there are $g^* \in P_{G^*}(f^*)$ and a primitive extremal signature σ of G^* such that

$$f^*(t) - g^*(t) = \sigma(t)\|f^* - g^*\| = \sigma(t)d(f^*, G^*), \qquad t \in \operatorname{supp}\sigma. \tag{5.58}$$

By Lemma 5.4, (5.58) implies that

$$\begin{aligned} \|f^* - p\| &\geqq \|f^* - p\|_{\operatorname{supp}\sigma} \\ &= \|f^* - g^* - (p - g^*)\|_{\operatorname{supp}\sigma} \\ &= \|d(f^*, G^*)\sigma - (p - g^*)\|_{\operatorname{supp}\sigma} \\ &\geqq \|d(f^*, G^*)\sigma\|_{\operatorname{supp}\sigma} + (\beta/\alpha)\|p - g^*\|_{\operatorname{supp}\sigma}, \qquad p \in G^*. \end{aligned} \tag{5.59}$$

Let $J=\{i\colon \operatorname{supp}\sigma\cap A_i\neq\varnothing\}$. Then for any $p\in G^*$,

$$\begin{aligned}\|p\|_{A(J)}&=\max\{\|p\|_{A_i}\colon i\in I\}\\ &\leqq\max\{\alpha\cdot\|p\|_{A_i\cap\operatorname{supp}\sigma}\colon i\in I\}\\ &\leqq\alpha\cdot\|p\|_{\operatorname{supp}\sigma}.\end{aligned}\tag{5.60}$$

It follows from (5.59) and (5.60) that

$$\|f^*-p\|\geqq d(f^*,G^*)+(\beta/\alpha^2)\cdot\|p-g^*\|_{A(J)},\qquad p\in G^*.\tag{5.61}$$

Formula (5.61) implies that $P_{G^*}(f^*)|_{A(J)}=\{g^*|_{A(J)}\}$. Thus, (5.61) is equivalent to

$$\|f^*-p\|\geqq d(f^*,G^*)+(\beta/\alpha^2)\cdot d(p|_{A(J)},P_{G^*}(f^*)|_{A(J)}),\qquad p\in G^*.\tag{5.62}$$

By (5.39) and (5.40),

$$d(p,G^*(A(J)))\leqq\gamma\cdot\|p\|_{A(J)},\qquad p\in G^*.\tag{5.63}$$

It follows from (5.56) that

$$\begin{aligned}\|h-p\|\geqq d(h|_{T^*\setminus A(J)},G_J)+\eta\cdot d(p|_{T^*\setminus A(J)},P_{G_J}(h)),\\ h\in C_0(T^*),\qquad p\in G_J.\end{aligned}\tag{5.64}$$

By (5.62)–(5.64) and Lemma 5.3, we get

$$\|f^*-g\|\geqq d(f^*,G^*)+\lambda^*d(g,P_{G^*}(f^*)),\qquad g\in G^*,\tag{5.65}$$

where

$$\lambda^*=\min\{\eta/4,\beta\cdot\eta/[\alpha^2\cdot\gamma\cdot 2(\eta+2)]\}.$$

If $f^*\in G^*$, then (5.65) is trivial. So (5.65) holds for any $f\in C_0(T)$. It follows from (5.65) and Lemma 5.2 that

$$\|f-p\|\geqq d(f,G)+\left(\frac{\lambda^*}{2}\right)\cdot d(p,P_G(f)),\qquad p\in G.$$

Thus, we have proved (5.57), since $\lambda=\lambda^*/2$. So, P_G is uniform Hausdorff strongly unique, i.e., (iv) implies (i) in Theorem 1.1.

REFERENCES

[1] M. BARTLETT, *On Lipschitz conditions, strong unicity and a theorem of A. K. Cline*, J. Approx. Theory, 14 (1975), pp. 245–250.

[2] V. I. BERDYSHEV, *Metric projection onto finite-dimensional subspaces of C and L*, Math. Zametki, 18 (1975), pp. 473–488. (In Russian.)

[3] J. BLATTER, P. D. MORRIS, AND D. E. WULBERT, *Continuity of the set-valued metric projection*, Math. Ann., 178 (1968), pp. 12–24.

[4] A. L. BROWN, *On continuous selections for metric projections in spaces of continuous functions*, J. Funct. Anal., 8 (1971), pp. 431–449.

[5] E. W. CHENEY, *Multivariate Approximation Theory: Selected Topics*, CBMS–NSF Regional Conference Series in Applied Mathematics 51, Society for Industrial and Applied Mathematics, Philadelphia, 1986.

[6] A. K. CLINE, *Lipschitz conditions on uniform approximation operators*, J. Approx. Theory, 8 (1973), pp. 160–172.

[7] F. DEUTSCH, *Linear selections for the metric projection*, J. Funct. Anal., 99 (1982), pp. 269–292.

[8] ———, *An exposition of recent results on continuous metric selections*, in Numerical Methods of Approximation Theory, L. Collatz, G. Meinardus, and G. Nürnberger, eds., Internat. Ser. Numer. Math., 81, Birkhäuser, Basel, Boston, 1987, pp. 67–79.

[9] F. DEUTSCH AND WU LI, *Strong uniqueness and Lipschitz continuity of metric projections in L_1*, preprint, 1989.

[10] F. DEUTSCH, WU LI, AND SUNG-HO PARK, *Characterization of continuous and Lipschitz continuous selections in normed linear spaces*, J. Approx. Theory, to appear.

[11] G. GODINI, *Best approximation and intersections of balls*, in Banach Space Theory and Its Applications, A. Pietsch, N. Popa, and I. Singer, eds., Lecture Notes in Math. 991, Springer-Verlag, Berlin, New York, 1983, pp. 44-54.

[12] T. FISCHER, *A continuity condition for the existence of a continuous selection for a set-valued mapping*, J. Approx. Theory, 49 (1987), pp. 340-345.

[13] WU LI, *Strong uniqueness and Lipschitz continuity of metric projections—a generalization of the classical Haar theory*, J. Approx. Theory, 56 (1989), pp. 164-184.

[14] ———, *The characterization of continuous selections for metric projections in $C(X)$*, Sci. Sinica Ser. A, 4 (1988), pp. 254-264. (In Chinese.) Vol. XXXI, 9 (1988), pp. 1039-1052. (In English.)

[15] ———, *Problems about continuous selections in $C(X)$ (IV): Characteristic description*, Acta Math. Sinica, 31 (1988), pp. 299-308.

[16] ———, *The intrinsic characterization of lower semicontinuity of metric projections in $C_0(T, X)$*, J. Approx. Theory, 57 (1989), pp. 136-149.

[17] ———, *Continuous selections of metric projections and regular weakly interpolating subspaces*, preprint, 1987.

[18] ———, *Continuous metric selection and multivariate approximation*, J. Math. Anal. Appl., to appear.

[19] W. A. LIGHT AND E. W. CHENEY, *Approximation Theory in Tensor Product Spaces*, Lecture Notes in Math. 1169, Springer-Verlag, Berlin, New York, 1985.

[20] P.-K. LIN, *Remarks on linear selections for the metric projection*, J. Approx. Theory, 43 (1985), pp. 64-74.

[21] D. J. NEWMAN AND H. S. SHAPIRO, *Some theorems on Chebyshev approximation*, Duke Math. J., 30 (1963), pp. 673-681.

[22] G. NÜRNBERGER, *Strong unicity constants in Chebyshev approximation*, Numerical Methods of Approximation Theory, Vol. 8, Internat. Ser. Numer. Math. 81, Birkhäuser, Basel, Boston, 1987, pp. 144-168.

[23] G. NÜRNBERGER AND M. SOMMER, *Continuous selections in Chebyshev approximation*, in Parametric Optimization and Approximation, Internat. Ser. Numer. Math. 72, Birkhäuser, Basel, Boston, 1984, pp. 248-263.

[24] ———, *A Remez type algorithm for spline functions*, Numer. Math., 41 (1983), pp. 117-146.

[25] S.-H. PARK, *Lipschitz continuous metric projections and selections*, Ph.D. thesis, Dept. of Mathematics, Pennsylvania State University, University Park, PA, 1987.

[26] ———, *Uniform Hausdorff strong uniqueness*, J. Approx. Theory, 58 (1989), pp. 78-89.

[27] K. PRZESLAWSKI, *Linear and Lipschitz continuous selectors for the family of convex sets in Euclidean vector spaces*, Bull. Acad. Polon. Sci. Ser. Sci. Tech., 33 (1985), pp. 31-33.

[28] J. R. RESPESS AND E. W. CHENEY, *On lipschitzian proximity maps*, in Nonlinear Analysis and Application, S. P. Singh and J. H. Burry, eds., Lecture Notes in Pure and Appl. Math. 80, Marcel Dekker, New York, 1982, pp. 73-85.

[29] T. J. RIVLIN AND H. S. SHAPIRO, *A unified approach to certain problems of approximation and minimization*, SIAM J. Appl. Math., 9 (1960), pp. 670-699.

[30] M. SOMMER, *Continuous selections and convergence of best L_p-approximation in subspaces of spline functions*, Numer. Funct. Anal. Optim., 6 (1983), pp. 213-234.

[31] V. STOVER, *The strict approximation and continuous selections for the metric projection*, Ph.D. thesis, University of California, San Diego, 1981.

[32] H. STRAUSS, *An algorithm for the computation of strict approximations in subspaces of spline functions*, J. Approx. Theory, 41 (1984), pp. 329-344.

[33] D. T. YOST, *Best approximation and intersections of balls in Banach spaces*, Bull. Austral. Math. Soc., 20 (1979), pp. 285-300.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 221-240, January 1990
© 1990 Society for Industrial and Applied Mathematics
012

JUSTIFICATION OF MATCHING WITH THE TRANSITION EXPANSION OF VAN DER POL'S EQUATION*

A. D. MacGILLIVRAY†

Abstract. The analysis of the relaxation oscillations of Van der Pol's equation presents an especially challenging test of the formal techniques of the method of matched asymptotic expansions for solving singular perturbation problems. The formal analysis is described in Kevorkian and Cole's monograph [J. Kevorkian and J. D. Cole, "Perturbation Methods in Applied Mathematics", Springer-Verlag, Berlin, New York, 1981], which explains why the inner and outer expansions must necessarily be supplemented by a third "transition" expansion in order to obtain a uniformly valid approximation beyond $O(1)$ on a complete half-period. Kevorkian and Cole carry out the construction and delicate matching of several terms in the expansions. The present paper mathematically justifies their formal results to $O(\varepsilon^{1/3})$, and is the first such proof for any transition expansion. Partly for this reason, but also because the idea underlying the proof has been and will be applied to other singular perturbation problems, this paper is intended to be a contribution to the study of asymptotic methods rather than merely to the theory of Van der Pol's equation.

Key words. singular perturbation, rigorous matching, Van der Pol, matched asymptotic expansions

AMS(MOS) subject classifications. 34C15, 34E15

1. Introduction.

This paper presents a mathematical justification of terms in the transition asymptotic expansion of relaxation oscillations of Van der Pol's equation, written as

$$\varepsilon \frac{d^2y}{dt^2} + (1-y^2)\frac{dy}{dt} + y = 0.$$

It also justifies matching these terms with the leading terms in the inner and outer asymptotic expansions. The construction of these terms and their formal (i.e., non-rigorous) matching is explained in complete detail in Kevorkian and Cole's well-known monograph [4].

Van der Pol's equation is, of course, a canonical example that has long been used in texts, exposition, and research. As an example illustrating the method of matched asymptotic expansions in solving singular perturbation problems, it provides an especially stringent test of that method because three principal asymptotic expansions are required, and the formal matching among them is an extremely delicate matter. It is not surprising that the mathematical justification also presents some challenges. Our analysis is believed to be the only justification of validity and matching of a transition asymptotic expansion for any nontrivial problem (and not merely the Van der Pol problem).

The present work extends the work of MacGillivray [8], [9], which presents a complete mathematical justification of the formal result by demonstrating that the leading terms of the inner and outer expansions, as constructed by Kevorkian and Cole [4], give $O(1)$ approximations to the solution, as ε tends to zero, on explicit domains of uniform validity that overlap. The extension by the present analysis justifies the assertion that the first two terms of the transition expansion and the leading terms of the inner and outer expansions give $O(\varepsilon^{1/3})$ approximations to the solution, as $\varepsilon \to 0$, on explicit domains of validity that overlap; see Fig. 1.

The question of how to organize this paper was difficult. I wanted to describe certain aspects of what Littlewood called the "dramatic fine structure of solutions"

* Received by the editors July 20, 1987; accepted for publication (in revised form) February 8, 1989.
† Department of Mathematics, State University of New York, Buffalo, New York 14214-3093.

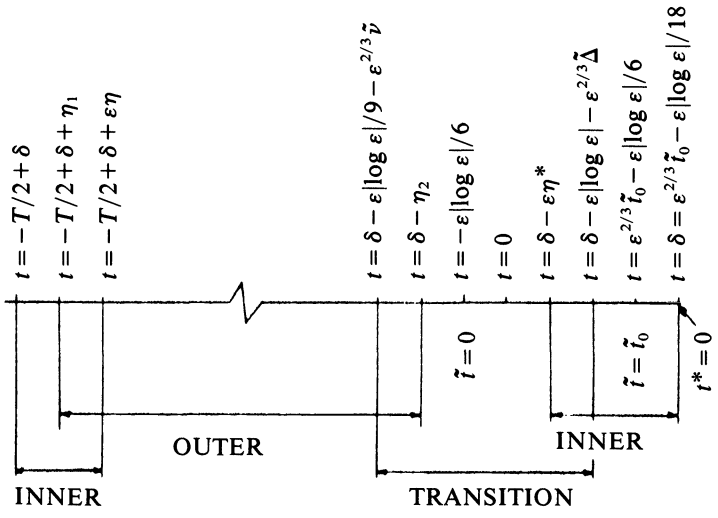


FIG. 1.

[7, p. 13], but sometimes detailed computations can get out of hand. (Littlewood describes his last paper on Van der Pol's equation as "The Monster" [7, p. 16]). However, the proofs include enough guideposts to enable the interested or skeptical reader to fill in some or all of the desired details. I hope I have achieved a balance between brevity and verbosity, but in any case, detailed proofs are available upon request from the author (MacGillivray [10], [11]).

The organization of the paper is as follows: Propositions 1–11 are presented in § 3, and lead to Theorem A, which is the first of four main theorems. Theorem A asserts that the first two terms of the transition expansion approximate the solution to $O(\varepsilon^{1/3})$ on an explicit domain of uniform validity. In § 4, the analysis begins with an application of Kaplun's Extension Theorem and, after five more propositions, concludes with Theorem B, which asserts an explicit domain of validity for the inner expansion to $O(\varepsilon^{1/3})$. The three propositions in § 5 lead to Theorem C, which gives an explicit domain of uniform validity of the outer expansion as an $O(\varepsilon^{1/3})$ approximation. In § 6, Theorem D collects these results together, completing the mathematical justification of the method of matched asymptotic expansions for this problem. It is worth mentioning that an interesting feature in the proofs is the recurring pairwise interaction between the expansions that is undoubtedly related in some way to the formal matching procedures.

The analysis of this and the two previous papers enables us to examine some formal techniques of asymptotic analysis from a different vantage point and thereby possibly enhance the understanding not only of how certain techniques work, but also why they work. This view has been highlighted by recent work on "fingering problems." Langer [6], for example, discusses one of these problems for which a straightforward perturbation analysis led to seemingly reasonable results. These were accepted as correct for many years, but Langer shows them to be incorrect. See also the review paper by Saffman [14].

For the sake of completeness, we give a short list of works devoted specifically to the analysis of Van der Pol's equation: Cartwright [1], Dorodnicyn [2], Haag [3], Pontryagin, Mishchenko, and coworkers; see [12], Stoker [15]. A recent paper by Storti and Rand [16] applies much of the formal analysis (including the transition expansion)

of Kevorkian and Cole to strongly coupled relaxation oscillators. Finally, we mention the book by O'Malley [13] and the review paper by Lagerstrom and Casten [5] to supplement Kevorkian and Cole's book [4] as references on singular perturbation problems and asymptotic expansions.

2. Notation and other preliminaries. In the next section we present the analysis that leads to Theorem A, which proves that the first two terms of the transition expansion constructed by Kevorkian and Cole [4] approximate the exact solution uniformly on an explicit interval as $\varepsilon \to 0$. We will use many results from Kevorkian and Cole's formal analysis as well as from MacGillivray [8], [9]. We assume the reader is familiar with the method of matched asymptotic expansions as developed in [4].

Recall Van der Pol's equation in the form

$$\varepsilon \frac{d^2y}{dt^2} + (1-y^2)\cdot\frac{dy}{dt} + y = 0. \tag{1}$$

We impose the same initial conditions as in [4]:

$$y(\delta) = 0, \tag{2}$$

$$y'(\delta) < 0 \tag{3}$$

where

$$\delta = \delta(\varepsilon) = \varepsilon^{2/3}\tilde{t}_0 - \frac{\varepsilon|\ln \varepsilon|}{18}, \tag{4}$$

with $\tilde{t}_0 = 2.3381\cdots$ being the absolute value of the first zero of the Airy function. The stretched inner variable t^* and transition variable $\tilde{t}$ are defined by the following expressions [4]:

$$\varepsilon t^* = t - \delta, \tag{5}$$

$$\varepsilon^{2/3}\tilde{t} = t + \frac{\varepsilon|\log \varepsilon|}{6}. \tag{6}$$

The relationships among t, t^*, and $\tilde{t}$ are shown schematically in Fig. 1.

We introduce the function $R(\tilde{t}; \varepsilon)$, and recall from [8], [9], the functions $r(t^*; \varepsilon)$ and $h(t; \varepsilon)$, all defined as corrections to terms in the asymptotic expansions constructed formally in [4]. They are

$$y(t; \varepsilon) = g_0(t^*) + r(t^*; \varepsilon), \tag{7}$$

$$y(t; \varepsilon) = u_0(t) + h(t; \varepsilon), \qquad t \leqq 0, \tag{8}$$

$$y(t; \varepsilon) = 1 + \varepsilon^{1/3} f_1(\tilde{t}) + \varepsilon^{1/3} R(\tilde{t}; \varepsilon), \qquad \tilde{t} < \tilde{t}_0. \tag{9}$$

$g_0(t^*)$ is defined in [4, form. 2.6.23), corrected:

$$\frac{1}{3}\ln(1-g_0) - \frac{1}{(1-g_0)} - \frac{1}{3}\ln(g_0+2) = t^*, \tag{10}$$

and $r(t^*; \varepsilon)$ satisfies (MacGillivray [8])

$$\frac{d^2r}{dt^{*2}} = \left(-\frac{dg_0}{dt^*}\right)\cdot(2g_0+r)r + (1-g_0^2-2rg_0-r^2)\frac{dr}{dt^*} - \varepsilon y. \tag{11}$$

$u_0(t)$ is defined in [4, form. (2.6.10)]:

$$\log u_0 - (u_0^2-1)/2 = t, \qquad t \leqq 0, \tag{12}$$

and $h(t;\varepsilon)$ satisfies (MacGillivray [9])

$$\varepsilon\frac{d^2h}{dt^2}=(1-y^2)\frac{dh}{dt}+\frac{(1+u_0y)h}{(u_0^2-1)}+\frac{\varepsilon u_0(1+u_0^2)}{(u_0^2-1)^3}. \tag{13}$$

Kevorkian and Cole construct $f_1(\tilde{t})$ using the Airy function. The reader is referred to their book for details, including their sketch of the graph of f_1 [4, Fig. 2.6.4]. Our Fig. 2 shows a careful plot of $-f_1(\tilde{t})$. $R(\tilde{t};\varepsilon)$ satisfies

$$\begin{aligned}\frac{d^2R}{d\tilde{t}^2}&=-2(R+f_1)\frac{dR}{d\tilde{t}}-2R\frac{df_1}{d\tilde{t}}-\varepsilon^{1/3}\\&\quad\cdot\left\{(f_1+R)+(f_1+R)^2\frac{dR}{d\tilde{t}}+(f_1+R)^2\frac{df_1}{d\tilde{t}}\right\}.\end{aligned} \tag{14}$$

It is proved in MacGillivray [9] that

$$y(t)=u_0+o(1)\quad\text{as }\varepsilon\to 0, \tag{15}$$

uniformly on any interval of the form

$$[-T/2+\mu_1(\varepsilon)+\delta,\ -\mu_2(\varepsilon)+\delta] \tag{16}$$

where

$$\varepsilon\ll\mu_i\ll 1,\qquad i=1,2 \tag{17}$$

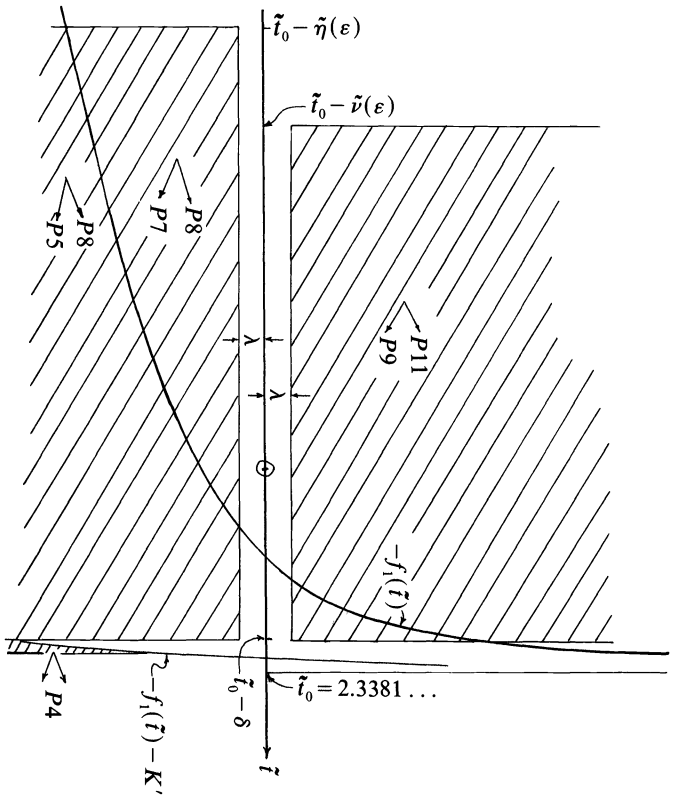


FIG. 2.

and where $T = T(\varepsilon)$, the period of oscillation, is calculated to be

$$T = 3 - 2\log 2 + o(1) \quad \text{as } \varepsilon \to 0 \tag{18}$$

in agreement with the well-known classical result.

MacGillivray [8] has proved that

$$y(t) = g_0(t^*) + O(\varepsilon^{1/3}) \quad \text{as } \varepsilon \to 0 \tag{19}$$

uniformly on the interval

$$t^* \in [-\varepsilon^{-1/3}/2,\ \varepsilon^{-2/3}]. \tag{20}$$

Stated otherwise, there exists a constant K such that for all sufficiently small ε and all t^* satisfying (20),

$$|y(t^*) - g_0(t^*)| < K\varepsilon^{1/3}. \tag{21}$$

Because $y(t;\ \varepsilon)$ is an odd periodic function, we have the further result

$$y(t) = -g_0(t^* + T/2\varepsilon) + O(\varepsilon^{1/3}) \tag{22}$$

uniformly on the interval

$$t^* \in \left[\frac{-T}{2\varepsilon} - \frac{\varepsilon^{-1/3}}{2},\ \frac{-T}{2\varepsilon} + \varepsilon^{-2/3}\right]. \tag{23}$$

The interval $t \in [-T/2 + \delta,\ \delta]$, is contained in the union of the three intervals (16), (20), (23) if ε is sufficiently small. Thus formulas (15), (19), and (22) together provide a complete $O(1)$ description of $y(t;\ \varepsilon)$ over a complete half-period, and hence for all t by odd periodic extension.

3. Analysis of the transition asymptotic expansion. The idea behind the following analysis is to construct regions in the $(y,\ t)$ plane that are "forbidden" to the solution. Consequently, the proofs often proceed by setting up contradiction arguments. These proofs have a strong geometric appeal, and so the reader may find Fig. 2 a useful guide. Notice the short line segments that appear in the crosshatched regions. The location of the tail of a line segment represents the point $(\tilde{t},\ R(\tilde{t}))$ and the sign of the slope of the line segment corresponds to the sign of $R'(\tilde{t})$. Beside each line segment is the number of the proposition that proves the corresponding $R(\tilde{t})$ and $R'(\tilde{t})$ is forbidden if ε is sufficiently small.

PROPOSITION 1. *Choose* $\alpha \in (0,\ \frac{1}{4}]$. *Then* $\Delta > 0$ *can be chosen to satisfy the following*:

(i) $\Delta < \frac{1}{2}$,

(ii) $(1+\alpha)/(\tilde{t} - \tilde{t}_0) < f_1(\tilde{t}) < (1-\alpha)/(\tilde{t} - \tilde{t}_0)$ *for all* $\tilde{t} \in [\tilde{t}_0 - \Delta,\ \tilde{t}_0]$,

(iii) $-(1+\alpha)/(\tilde{t} - \tilde{t}_0)^2 < f_1'(\tilde{t}) < -(1-\alpha)/(\tilde{t} - \tilde{t}_0)^2$ *for all* $\tilde{t} \in [\tilde{t}_0 - \Delta,\ \tilde{t}_0)$,

(iv) $1/\Delta > K$, *where* K *is defined in* (21).

Proof. Inequalities (ii), (iii) follow from [4, forms. (2.6.58); (2.6.59), corrected].

PROPOSITION 2. *Let* Δ *be as in Proposition* 1. *Then, for all sufficiently small* ε, *the following properties hold*:

(i) $R(\tilde{t}_0 - \Delta/2) > -f_1(\tilde{t}_0 - \Delta/2) - K'$ *for some constant* K',

(ii) $R(\tilde{t}_0 - \Delta) < -f_1(\tilde{t}_0 - \Delta)$ (*strict inequality*),

(iii) $R(\tilde{t}_0 - \varepsilon^{1/6}) = o(\varepsilon^{-1/6})$,

(iv) $R(\tilde{t}_0 - \varepsilon^{1/3}|\log \varepsilon|) \to \infty$ *as* $\varepsilon \to 0$,

(v) $R(\tilde{t}) + f_1(\tilde{t}) > -19\varepsilon^{-1/3}/(20|\log \varepsilon|)$ *for* $\tilde{t} \in [\tilde{t}_0 - \Delta,\ \tilde{t}_0 - \varepsilon^{1/3}|\log \varepsilon|]$.

Proof. Using (5), (6), (19), and (20), we can readily show

$$y(t;\ \varepsilon) = g_0\left((\tilde{t} - \tilde{t}_0)\varepsilon^{-1/3} - \frac{|\log \varepsilon|}{9}\right) + O(\varepsilon^{1/3}) \tag{24}$$

uniformly for $\tilde{t} \in [\tilde{t}_0 - \Delta, \tilde{t}_0)$. From [4, form. (2.6.24)],

$$g_0(t^*) = 1 + \frac{1}{t^*} - \frac{\log(-t^*)}{(3t^{*2})} + \cdots \quad \text{as } t^* \to -\infty, \tag{25}$$

so that (24) can be rewritten as

$$y(t;\varepsilon) = 1 + \frac{1+o(1)}{(\tilde{t}-\tilde{t}_0)\varepsilon^{-1/3} - \dfrac{|\log\varepsilon|}{9}} + O(\varepsilon^{1/3}). \tag{26}$$

Thus,

$$\varepsilon^{1/3}(f_1(\tilde{t}) + R(\tilde{t})) = \frac{\varepsilon^{1/3}(1+o(1))}{(\tilde{t}-\tilde{t}_0)\dfrac{1-\varepsilon^{1/3}|\log\varepsilon|}{9(\tilde{t}-\tilde{t}_0)}} + O(\varepsilon^{1/3}). \tag{27}$$

Recalling the definition of K in (21), we easily show that

$$-\frac{4}{\Delta} < f_1\left(\tilde{t}_0 - \frac{\Delta}{2}\right) + R\left(\tilde{t}_0 - \frac{\Delta}{2}\right) < 0, \tag{28}$$

$$-\frac{4}{\Delta} < f_1(\tilde{t}_0 - \Delta) + R(\tilde{t}_0 - \Delta) < 0, \tag{29}$$

and (i) and (ii) follow immediately, with $K' = 4/\Delta$.

To prove (iii), simply substitute $\tilde{t} = \tilde{t}_0 - \varepsilon^{1/6}$ into (26) and use the following result from [4]:

$$f_1(\tilde{t}) = \frac{1}{(\tilde{t}-\tilde{t}_0)} - \frac{\tilde{t}_0(\tilde{t}-\tilde{t}_0)}{3} + O((\tilde{t}-\tilde{t}_0)^2) \quad \text{as } \tilde{t} \to \tilde{t}_0. \tag{30}$$

Part (iv) is proved in a similar fashion. The proof of (v) also depends on (21) and (30) and is the result of a straightforward computation.

PROPOSITION 3. *For all sufficiently small ε, the amplitude $y(t)$ is positive on the interval $\tilde{t} \in [\tilde{t}_0 - \varepsilon^{-2/3}/2, \tilde{t}_0)$ and, in addition, $R(\tilde{t}_0 - \varepsilon^{-2/3}/2)$ is positive.*

Proof. $y(t)$ is positive between its zeros at $t = \delta - T/2$ and $t = \delta$. Writing this interval in terms of $\tilde{t}$ using (6), a short computation shows the interval $\tilde{t} \in [\tilde{t}_0 - \varepsilon^{-2/3}/2, \tilde{t}_0)$ lies within it, provided ε is sufficiently small. This verifies the first part of the proposition.

To prove the second part, first note that the point $\tilde{t} = \tilde{t}_0 - \varepsilon^{-2/3}/2$ corresponds to $t = -\frac{1}{2} + o(1)$, which is well within the domain of uniform validity of $u_0(t)$; see (15) and (16). Thus it follows that

$$1 + \varepsilon^{1/3}\left[f_1\left(\tilde{t}_0 - \frac{\varepsilon^{-2/3}}{2}\right) + R\left(\tilde{t}_0 - \frac{\varepsilon^{-2/3}}{2}\right)\right] = u_0\left(-\frac{1}{2}\right) + o(1). \tag{31}$$

From [4, form. (2.6.50)],

$$\begin{aligned} f_1\left(\tilde{t}_0 - \frac{\varepsilon^{-2/3}}{2}\right) &= \sqrt{\left(\frac{\varepsilon^{-2/3}}{2} - \tilde{t}_0\right)} + \cdots \\ &= \varepsilon^{-1/3}\left(\frac{1}{\sqrt{2+o(1)}}\right) \quad \text{as } \varepsilon \to 0. \end{aligned} \tag{32}$$

Thus

$$R\left(\tilde{t}_0-\frac{\varepsilon^{-2/3}}{2}\right)=\varepsilon^{-1/3}\left(-\frac{1}{\sqrt{2+o(1)}}+u_0\left(-\frac{1}{2}\right)-1\right). \tag{33}$$

It is a trivial matter to estimate $u_0(-\frac{1}{2})$ numerically from (12):

$$u_0(-\tfrac{1}{2})=1.7333\cdots. \tag{34}$$

Substituting (34) into (33) leads immediately to the verification of the second part of the proposition.

PROPOSITION 4. *Let Δ be chosen as in Proposition* 1, *and K' as in Proposition* 2. *Then for all sufficiently small ε,*

$$R(\tilde{t})\geqq -f_1(\tilde{t})-K' \quad \textit{for } \tilde{t}\in\left[\tilde{t}_0-\varepsilon^{-2/3},\ \tilde{t}_0-\frac{\Delta}{2}\right]. \tag{35}$$

Proof. It is easy to show, using part (i) of Proposition 2, and Proposition 3, that, if ε is sufficiently small, $R(\tilde{t})$ exceeds $-f_1(\tilde{t})-K'$ when $\tilde{t}=\tilde{t}_0-\Delta/2$ and when $\tilde{t}=\tilde{t}_0-\varepsilon^{-2/3}/2$. Making the tentative assumption that the conclusion of the proposition is false implies the existence of $\tilde{t}_1$, $\tilde{t}_2$, we have

$$\tilde{t}_0-\frac{\varepsilon^{-2/3}}{2}<\tilde{t}_1<\tilde{t}_2<\tilde{t}_0-\frac{\Delta}{2}$$

such that

$$R(\tilde{t}_1)+f_1(\tilde{t}_1)=R(\tilde{t}_2)+f_1(\tilde{t}_2)=-K',$$

$$R(\tilde{t})+f_1(\tilde{t})<-K' \quad \text{for } \tilde{t}\in(\tilde{t}_1,\tilde{t}_2).$$

This leads immediately to a $\tilde{t}_3\in(\tilde{t}_1,\tilde{t}_2)$, where y lies between zero and unity and where y' is zero. Such a point lies on a phase plane trajectory that spirals toward the origin as $\tilde{t}\to-\infty$, and is therefore not on the limit cycle. This contradiction completes the proof.

PROPOSITION 5. *Let $1\ll\tilde{\eta}(\varepsilon)\ll\varepsilon^{-2/3}/2$, let $\lambda>0$ be given, and let Δ be as in Proposition* 1. *Then, for all sufficiently small ε, there exists no $\tilde{t}_1\in[\tilde{t}_0-\tilde{\eta},\ \tilde{t}_0-\Delta]$ for which $R(\tilde{t}_1)\leqq-\lambda$, $R(\tilde{t}_1)\leqq-f_1(\tilde{t}_1)$, and $R'(\tilde{t}_1)\leqq 0$.*

Proof. Assume tentatively that such a $\tilde{t}_1$ exists, and note that for any $\tilde{t}\in[\tilde{t}_0-\tilde{\eta},\ \tilde{t}-\Delta/2]$ for which $R(\tilde{t})\leqq-\lambda$, $R(\tilde{t})\leqq-f_1(\tilde{t})$, and $R'(\tilde{t})\leqq 0$, Proposition 4 and (14) imply

$$\frac{d^2R}{d\tilde{t}^2}\leqq 2\lambda\left(\frac{df_1}{d\tilde{t}}\right)-\varepsilon^{1/3}(f_1+R)^2\left(\frac{dR}{d\tilde{t}}\right)+o(\tilde{\eta}^{-1/2}) \tag{36}$$

for all sufficiently small ε.

From the text above equation (2.6.50) in [4], we find $f_1'(\tilde{t})=\sqrt{(-\tilde{t})}+\cdots$ as $\tilde{t}\to-\infty$. Since $|f_1'|$ is an increasing function, it follows directly that

$$\left|\frac{df_1}{d\tilde{t}}\right|\geqq\left(\frac{\tilde{\eta}^{-1/2}}{2}\right)(1+o(1)) \quad \text{for } \tilde{t}\in[\tilde{t}_0-\tilde{\eta},\ \tilde{t}_0).$$

Thus

$$\frac{d^2R}{d\tilde{t}^2}\leqq-2\lambda\left(\frac{\tilde{\eta}^{-1/2}}{2}\right)(1+o(1))-\varepsilon^{1/3}(f_1+R)^2\left(\frac{dR}{d\tilde{t}}\right), \tag{37}$$

assuming the conditions above (36) hold. It is easy now to use a continuation argument to verify that on the interval $(\tilde{t}_1, \tilde{t}_0-\Delta/2)$ we have

$$R'(\tilde{t})<0, \quad R(\tilde{t})<-\lambda, \quad R(\tilde{t})<-f_1(\tilde{t}).$$

On the interval $[\tilde{t}_0-\Delta, \tilde{t}_0-\Delta/2]$, $R(\tilde{t})$ is bounded. Thus, if ε is sufficiently small, the two terms containing $R'(\tilde{t})$ in (14) combine to give a negative contribution. The remaining terms in the braces being bounded yields

$$\frac{d^2R}{d\tilde{t}^2}\leqq -2\lambda|f_1'(\tilde{t}_0-\Delta)|+o(1) \tag{38}$$

on the interval $[\tilde{t}_0-\Delta, \tilde{t}_0-\Delta/2]$. Two integrations yield

$$R'(\tilde{t})<-2\lambda|f_1'(\tilde{t}_0-\Delta)|\cdot(1+o(1))\cdot(\tilde{t}-\tilde{t}_0+\Delta), \tag{39}$$

$$R(\tilde{t})<-\lambda-2\lambda|f_1'(\tilde{t}_0-\Delta)|\cdot(1+o(1))\cdot\frac{(\tilde{t}-\tilde{t}_0+\Delta)^2}{2} \tag{40}$$

on $(\tilde{t}_0-\Delta, \tilde{t}_0-\Delta/2]$.

Let

$$\kappa=\left(\frac{3\lambda\Delta^2}{32}\right)|f_1'(\tilde{t}_0-\Delta)|. \tag{41}$$

Then (39), (40) give

$$R\left(\tilde{t}_0-\frac{\Delta}{2}\right)<\frac{\kappa}{\left(\tilde{t}_0-\frac{\Delta}{2}\right)-\tilde{t}_0}, \tag{42}$$

$$R'\left(\tilde{t}_0-\frac{\Delta}{2}\right)<\frac{-\kappa}{\left[\left(\tilde{t}_0-\frac{\Delta}{2}\right)-\tilde{t}_0\right]^2}. \tag{43}$$

Now as long as

$$R(\tilde{t})\leqq\frac{\kappa}{\tilde{t}-\tilde{t}_0}, \tag{44}$$

$$R'(\tilde{t})\leqq\frac{-\kappa}{(\tilde{t}-\tilde{t}_0)^2} \tag{45}$$

remain valid to the right of $\tilde{t}=\tilde{t}_0-\Delta/2$, both (f_1+R) and R' will remain negative. Then estimate (v) in Proposition 2, together with properties (ii), (iii) in Proposition 1, leads, after a straightforward computation, to the conclusion that as long as (44) and (45) remain valid to the right of $\tilde{t}_0-\Delta/2$, $\tilde{t}<\tilde{t}_0$,

$$\frac{d^2R}{d\tilde{t}^2}\leqq\left[\frac{2\kappa}{(\tilde{t}-\tilde{t}_0)^3}\right]\{(2-2\alpha+\kappa)\cdot(1+o(1))\}+o(1), \tag{46}$$

which is negative if ε is sufficiently small. Furthermore, the quantity within the braces exceeds unity if ε is sufficiently small, and an obvious continuation argument leads to the conclusion that (44), (45) remain valid on $[\tilde{t}_0-\Delta/2, \tilde{t}_0-\varepsilon^{1/3}|\log\varepsilon|]$. This contradicts (iv) of Proposition 2, if ε is sufficiently small, completing the proof.

COROLLARY. *The conclusion of Proposition* 5 *remains true if* $\tilde{\eta}(\varepsilon)$ *is replaced by* $\tilde{\eta}(\varepsilon)+1$.

PROPOSITION 6. *If ε is sufficiently small, $R(\tilde{t})$ cannot have a minimum in the region bounded by*

$$-f_1(\tilde{t}) \leqq R(\tilde{t}) \leqq -\lambda \quad \text{and} \quad \tilde{t}_0 - \varepsilon^{-2/3} \leqq \tilde{t} \leqq \tilde{t}_0 - \Delta.$$

Proof. If we assume tentatively the existence of such a minimum at $\tilde{t}_1$, then at $\tilde{t}_1$ we show easily that

$$\frac{d^2R}{d\tilde{t}^2} \leqq -2R\left(\frac{df_1}{d\tilde{t}}\right) - \varepsilon^{1/3}[(f_1+R)+(f_1+R)f_1(\tilde{t}_1)\cdot f_1'(\tilde{t}_1)]. \tag{47}$$

From what precedes equation (2.6.50) in [4], we find

$$f_1(\tilde{t}) = \sqrt{(-\tilde{t})} + \cdots \quad \text{and} \quad f_1'(\tilde{t}) = \frac{-1}{2\sqrt{(-\tilde{t})}} + \cdots \tag{48}$$

as $\tilde{t} \to -\infty$. Then there exists a $\tilde{\tau}$ such that for $\tilde{t} \leqq \tilde{t}_0 - \tilde{\tau}$,

$$f_1(\tilde{t}) < (1+\alpha)\sqrt{(-\tilde{t})} \quad \text{and} \quad f_1'(\tilde{t}) > -(1+\alpha)[2\sqrt{(-\tilde{t})}]. \tag{49}$$

Two easy computations (one if $\tilde{t}_1 < \tilde{t}_0 - \tilde{\tau}$, and the other if $\tilde{t}_1 \in [\tilde{t}_0 - \tilde{\tau}, \tilde{t}_0 - \Delta]$) yield $R''(\tilde{t}_1) < 0$ if ε is sufficiently small, a contradiction.

PROPOSITION 7. *Let $\tilde{\eta}$, λ, and Δ be as in Proposition 5. Then, for all sufficiently small ε, there exists no $\tilde{t}_1 \in [\tilde{t}_0 - \tilde{\eta}, \tilde{t}_0 - \Delta]$ for which*

$$-f_1(\tilde{t}_1) \leqq R(\tilde{t}_1) < -\lambda \quad \text{and} \quad R'(\tilde{t}_1) \leqq 0. \tag{50}$$

Proof. Assume tentatively such a $\tilde{t}_1$ exists. An obvious continuation argument leads to the conclusion that as long as $R(\tilde{t}) \geqq -f_1(\tilde{t})$ to the right of $\tilde{t}_1$, $R'(\tilde{t})$ will be negative and lie below $-\lambda$. Clearly (see Fig. 2) this cannot persist indefinitely, and the graph of $R(\tilde{t})$ must enter the region $R \leqq -f_1(\tilde{t})$, $R \leqq -\lambda$, $\tilde{t}_0 - \tilde{\eta} \leqq \tilde{t} \leqq \tilde{t}_0 - \Delta$ with a nonpositive slope. Proposition 5 then yields the desired contradiction.

COROLLARY. *The conclusion of Proposition 7 remains true if $\tilde{\eta}$ is replaced by $\tilde{\eta}+1$.*

PROPOSITION 8. *Let $\tilde{\eta}$, λ, and Δ be as in Proposition 5 Then, if ε is sufficiently small, there exists no $\tilde{t}_2 \in [\tilde{t}_0 - \tilde{\eta}, \tilde{t}_0 - \Delta]$ for which*

$$R'(\tilde{t}_2) > 0 \quad \text{and} \quad R(\tilde{t}_2) \leqq -\lambda. \tag{51}$$

Proof. Assume tentatively such a $\tilde{t}_2$ exists. It follows from the corollary to Proposition 5 and the proof of Proposition 6 that

$$R'(\tilde{t}_0 - \tilde{\eta}) > 0 \quad \text{and} \quad R(\tilde{t}_0 - \tilde{\eta}) \leqq -\lambda. \tag{52}$$

There are two cases, depending on whether $R(\tilde{t}_0 - \tilde{\eta}) \geqq -f_1(\tilde{t}_0 - \tilde{\eta})$ or not. If $R(\tilde{t}_0 - \tilde{\eta}) \geqq -f_1(\tilde{t}_0 - \tilde{\eta})$, then, as $\tilde{t}$ decreases, $R(\tilde{t})$ cannot drop below $-f_1(\tilde{t})$ on the interval $[\tilde{t}_0 - \tilde{\eta} - 1, \tilde{t}_0 - \tilde{\eta}]$ and still be on the limit cycle (recall $R(\tilde{t}_0 - \varepsilon^{-2/3}/2)$ is positive) and so

$$R(\tilde{t}_0 - \tilde{\eta} - 1) \geqq -f_1(\tilde{t}_0 - \tilde{\eta} - 1). \tag{53}$$

The same conclusion holds if the other alternative, $R(\tilde{t}_0 - \tilde{\eta}) < -f_1(\tilde{t}_0 - \tilde{\eta})$, is assumed. To show this, note that $R'(\tilde{t}_0 - \tilde{\eta}) < -f_1'(\tilde{t}_0 - \tilde{\eta})$ if ε is sufficiently small, since otherwise we would not be on the limit cycle. Then as long as

$$R(\tilde{t}) \leqq -f_1(\tilde{t}) \quad \text{and} \quad R'(\tilde{t}) \leqq -f_1'(\tilde{t}) \tag{54}$$

to the left of $\tilde{t}_0 - \tilde{\eta}$, we show with the help of Proposition 4 that

$$\frac{d^2R}{d\tilde{t}^2} \leqq 2f_1\left(\frac{df_1}{d\tilde{t}}\right) + O(\varepsilon^{1/3}), \tag{55}$$

and estimates (48) then give

$$\frac{d^2R}{d\tilde{t}^2} \leqq \{2\sqrt{(\tilde{t}_0-\tilde{\eta})(1+o(1))}\}\left\{\frac{-1}{2\sqrt{(\tilde{t}_0-\tilde{\eta})}}(1+o(1))\right\}+O(\varepsilon^{1/3}). \tag{56}$$

Thus if ε is sufficiently small,

$$\frac{d^2R}{d\tilde{t}^2} \leqq -\frac{1}{3} \tag{57}$$

as long as (54) holds to the left of $\tilde{t}_0-\tilde{\eta}$. This implies

$$R'(\tilde{t}_0-\tilde{\eta}-1) > -f_1'(\tilde{t}_0-\tilde{\eta}-1) \tag{58}$$

if ε is sufficiently small, which in turn implies (53) (for otherwise the trajectory is not the limit cycle). At this stage it has been established that for ε sufficiently small,

$$-f_1(\tilde{t}_0-\tilde{\eta}-1) \leqq R(\tilde{t}_0-\tilde{\eta}-1) < -\lambda, \tag{59a}$$

$$R'(\tilde{t}_0-\tilde{\eta}-1) > 0. \tag{59b}$$

Now, to be on the limit cycle trajectory, $R(\tilde{t}) \geqq -f_1(\tilde{t})$ for all $\tilde{t} \in [\tilde{t}_0-\varepsilon^{-2/3}/2,\ \tilde{t}_0-\tilde{\eta}-1]$. However, from the proof of Proposition 6, $R'(\tilde{t})$ remains positive to the left of $\tilde{t}_0-\tilde{\eta}-1$ at least until $\tilde{t}_0-\varepsilon^{-2/3}/2$ is reached, forcing $R(\tilde{t}_0-\varepsilon^{-2/3}/2) < -\lambda$. This contradicts Proposition 3.

So far it has been shown that if ε is sufficiently small, no part of the graph of $R(\tilde{t})$ can appear in the region $R \leqq -\lambda$, $\tilde{t}_0-\tilde{\eta} \leqq \tilde{t} \leqq \tilde{t}_0-\Delta$. The next three propositions explore the region $R \geqq \lambda$.

PROPOSITION 9. *Let λ, Δ be as in Proposition 5, and let $\tilde{\nu} = \tilde{\nu}(\varepsilon)$ satisfy $\Delta \ll \tilde{\nu} \ll \varepsilon^{-1/3}$. Then, for all sufficiently small ε, there exists no $\tilde{t}_1 \in [\tilde{t}_0-\tilde{\nu},\ \tilde{t}_0-\Delta]$ for which*

$$R(\tilde{t}_1) \geqq \lambda \quad \textit{and} \quad R'(\tilde{t}_1) \leqq 0. \tag{60}$$

Proof. Assume tentatively that such a $\tilde{t}_1$ exists, and note that wherever R' is nonpositive, (14) gives

$$\frac{d^2R}{d\tilde{t}^2} \geqq -2(R+f_1)\left(\frac{dR}{d\tilde{t}}\right)-2R\left(\frac{df_1}{d\tilde{t}}\right)-\varepsilon^{1/3}(f_1+R). \tag{61}$$

The idea of the proof is to show R' is negative for all $\tilde{t} < \tilde{t}_1$; this would contradict the fact that y is periodic. We begin by showing $d^2R/d\tilde{t}^2$ is positive on the interval $[\tilde{t}_0-2\tilde{\nu},\ \tilde{t}_1)$. Since the first term on the right of (61) has factor $dR/d\tilde{t}$, we need only show that the second factor dominates the third. Specifically, we can easily show that at any $\tilde{t}$ in the interval $[\tilde{t}_0-2\tilde{\nu},\ \tilde{t}_1)$ at which $R(\tilde{t}) \geqq \lambda$,

$$|\varepsilon^{1/3}(f_1+R)| < \frac{1}{2}\left|2R\left(\frac{df_1}{d\tilde{t}}\right)\right| \tag{62}$$

for all sufficiently small ε. Having shown this dominance, an obvious continuation argument leads to the conclusion that for all sufficiently small ε,

$$\frac{dR}{d\tilde{t}} < 0 \quad \text{and} \quad R(\tilde{t}) > \lambda \quad \text{on } [\tilde{t}_0-2\tilde{\nu},\ \tilde{t}_1). \tag{63}$$

If ε sufficiently small,

$$f_1(\tilde{t})+R(\tilde{t}) > 0 \quad \text{and} \quad f'(\tilde{t}) \leqq -\frac{3}{8\sqrt{(\tilde{t}_0-\tilde{t})}} \quad \text{on } [\tilde{t}_0-2\tilde{\nu},\ \tilde{t}_0-\tilde{\nu}] \tag{64}$$

so that

$$\frac{d^2R}{d\tilde{t}^2} > \left(\frac{3\lambda}{8}\right)(\tilde{t}_0 - \tilde{t})^{-1/2}, \qquad \tilde{t} \in [\tilde{t}_0 - 2\tilde{\nu}, \tilde{t}_0 - \tilde{\nu}], \tag{65}$$

and an integration gives

$$R'(\tilde{t}_0 - 2\tilde{\nu}) \leqq -\left(\frac{3\lambda}{4}\right)\left(1 - \frac{1}{\sqrt{2}}\right)(2\tilde{\nu})^{1/2}. \tag{66}$$

Thus the first term in (61) dominates the third term when $\tilde{t} = \tilde{t}_0 - 2\tilde{\nu}$. Again an easy continuation argument can be applied to the left of $\tilde{t}_0 - 2\tilde{\nu}$ to conclude $R'(\tilde{t})$ is negative for all $\tilde{t} < \tilde{t}_1$, which implies the contradiction mentioned near the beginning of the proof.

PROPOSITION 10. *Let $\tilde{\nu} = \tilde{\nu}(\varepsilon)$ again satisfy $\Delta \ll \tilde{\nu} \ll \varepsilon^{-1/3}$ with Δ, λ as in Proposition 5. If there is a $\tilde{t}_1 \in [\tilde{t}_0 - \tilde{\nu}, \tilde{t}_0 - \Delta]$ for which $R(\tilde{t}_1) \geqq \lambda$ and $R'(\tilde{t}_1) > 0$, then, provided ε is sufficiently small, $R(\tilde{t}) > \lambda$ and $R'(\tilde{t}) > 0$ for all $\tilde{t} \in (\tilde{t}_1, \tilde{t}_0) \subset [\tilde{t}_0 - \tilde{\nu}, \tilde{t}_0)$. In addition there exists λ' which depends on λ and Δ but is independent of sufficiently small ε such that*

$$R'\left(\tilde{t}_0 - \frac{\Delta}{2}\right) \geqq \lambda'. \tag{67}$$

Proof. Assume such a $\tilde{t}_1$ exists. From Proposition 9 we conclude immediately that $R(\tilde{t}_0 - \Delta) > \lambda$ and $R'(\tilde{t}_0 - \Delta) > 0$. To prove these inequalities on $[\tilde{t}_0 - \Delta, \tilde{t}_0)$ note that from (ii) of Proposition 2, $R'(\tilde{t}_0 - \Delta) < -f_1(\tilde{t}_0 - \Delta)$. From this it follows that

$$R(\tilde{t}) < -f_1(\tilde{t}) \quad \text{and} \quad R'(\tilde{t}) < -f_1'(\tilde{t}) \quad \text{on } [\tilde{t}_0 - \Delta, \tilde{t}_0), \tag{68}$$

for otherwise the trajectory is not the limit cycle. The information above used in (14) leads directly to

$$\frac{d^2R}{d\tilde{t}^2} > -2R\left(\frac{df_1}{d\tilde{t}}\right) \quad \text{on } [\tilde{t}_0 - \Delta, \tilde{t}_0), \tag{69}$$

from which the conclusion $R(\tilde{t}) > \lambda$, $R'(\tilde{t}) > 0$ on $(\tilde{t}_1, \tilde{t}_0)$ follows immediately. To complete the proof simply integrate (69) from $\tilde{t}_0 - \Delta$ to $\tilde{t}_0 - \Delta/2$. The result is

$$R'\left(\tilde{t}_0 - \frac{\Delta}{2}\right) > \lambda |f_1'(\tilde{t}_0 - \Delta)| \left(\frac{\Delta}{2}\right) \equiv \lambda'. \tag{70}$$

The next proposition proves that $\tilde{t}_1$ as described in Proposition 10 cannot exist.

PROPOSITION 11. *Let $\tilde{\nu}$ be as in Proposition 10, and let Δ and λ be as in Proposition 5. Let λ' be defined by (70). Then, for all sufficiently small ε, there is no $\tilde{t}_1 \in [\tilde{t}_0 - \tilde{\nu}, \tilde{t}_0 - \Delta]$ for which $R(\tilde{t}_1) \geqq \lambda$ and $R'(\tilde{t}_1) > 0$.*

Proof. Assume tentatively that such a $\tilde{t}_1$ exists. From Proposition 10,

$$R\left(\tilde{t}_0 - \frac{\Delta}{2}\right) > \lambda \quad \text{and} \quad R'\left(\tilde{t}_0 - \frac{\Delta}{2}\right) > \lambda' \tag{71}$$

for all sufficiently small ε. From (71) there exists κ_0, independent of ε, such that

$$\kappa_0 R\left(\tilde{t}_0 - \frac{\Delta}{2}\right) > -f_1\left(\tilde{t}_0 - \frac{\Delta}{2}\right) - \frac{1}{2f_1'\left(\tilde{t}_0 - \frac{\Delta}{2}\right)}, \tag{72}$$

$$\kappa_0 R'\left(\tilde{t}_0 - \frac{\Delta}{2}\right) \geqq -f_1'\left(\tilde{t}_0 - \frac{\Delta}{2}\right) \tag{73}$$

for all sufficiently small ε. Now multiply (69) by κ_0:

$$\frac{d^2(\kappa_0 R)}{d\tilde{t}^2} > -2(\kappa_0 R)\left(\frac{df_1}{d\tilde{t}}\right), \qquad \tilde{t} \in \left[\tilde{t}_0 - \frac{\Delta}{2}, \tilde{t}_0\right) \tag{74}$$

and add it to the equation satisfied by f_1 [4, p. 75]. The result is the inequality

$$\frac{d^2(\kappa_0 R + f_1)}{d\tilde{t}^2} > -2(\kappa_0 R + f_1)\left(\frac{df_1}{d\tilde{t}}\right) - 1, \qquad \tilde{t} \in \left[\tilde{t}_0 - \frac{\Delta}{2}, \tilde{t}_0\right). \tag{75}$$

Inequalities (72) and (73) imply that at $\tilde{t}_0 - \Delta/2$, $d^2(\kappa_0 R + f_1)/d\tilde{t}^2$ is positive and that $d(\kappa_0 R + f_1)/d\tilde{t}$ is positive on some interval to the right of $\tilde{t}_0 - \Delta/2$, so $\kappa_0 R + f_1$ is increasing. f_1', on the other hand, is negative and decreasing. A continuation argument then yields

$$\kappa_0 R(\tilde{t}) + f_1(\tilde{t}) > 0 \quad \text{for all } \tilde{t} \in \left[\tilde{t}_0 - \frac{\Delta}{2}, \tilde{t}_0\right). \tag{76}$$

In particular,

$$R(\tilde{t} - \varepsilon^{1/6}) > \frac{\varepsilon^{-1/6}(1 + o(1))}{\kappa_0} \tag{77}$$

where (30) has been used. But (77) contradicts (iii) of Proposition 2, so $\tilde{t}_1$ cannot exist if ε is sufficiently small.

The first main result can now be proved.

THEOREM A. *Let Δ be any positive number and let $\tilde{\nu} = \tilde{\nu}(\varepsilon)$ satisfy*

$$\Delta \ll \tilde{\nu}(\varepsilon) \ll \varepsilon^{-1/3}. \tag{78}$$

Then

$$y(t) = 1 + \varepsilon^{1/3} f_1(\tilde{t}) + o(\varepsilon^{1/3}) \tag{79}$$

uniformly for

$$\tilde{t} \in [\tilde{t}_0 - \tilde{\nu}, \tilde{t}_0 - \Delta]. \tag{80}$$

Proof. Let λ be any positive number, and note that, with no loss of generality, we can assume Δ satisfies the conditions in Proposition 1. Note also that $\tilde{\nu} \ll \varepsilon^{-1/3} \ll \varepsilon^{-2/3}$, and so $\tilde{\nu}$ can replace $\tilde{\eta}$. From Propositions 5, 7-9, and 11,

$$|R(\tilde{t})| < \lambda \quad \text{for } \tilde{t} \in [\tilde{t}_0 - \tilde{\nu}, \tilde{t}_0 - \Delta] \tag{81}$$

for all sufficiently small ε. That is,

$$R(\tilde{t}) = o(1) \quad \text{for } \tilde{t} \in [\tilde{t}_0 - \tilde{\nu}, \tilde{t}_0 - \Delta] \tag{82}$$

and from (9) the result follows.

4. Inner expansion analysis. We begin with a summary of results proved in [8] concerning the leading term of the inner asymptotic expansion. Recall $r(t^*)$ is defined in (7).

PROPOSITION 12.

(a) $|y(t; \varepsilon)| < 3$ *if* $\varepsilon < \frac{1}{15}$,

(b) $dr/dt^*(0) < 0$,

(c) $r(t^*) \geqq 0$ *for* $t^* \in [-\varepsilon^{-1/3}/2, 0]$,

(d) $y(t; \varepsilon) = g_0(t^*) + O(\varepsilon^{1/3})$,

uniformly for $t^* \in [-\varepsilon^{-1/3}/2, +\varepsilon^{-2/3}]$.

The next proposition extends the domain of uniform validity of the transition expansion declared in Theorem A, and is an immediate consequence of Kaplun's Extension Theorem [4], [5].

PROPOSITION 13. *Let* $\tilde{\nu} = \tilde{\nu}(\varepsilon)$ *be as in Theorem* A. *Then there exists* $\tilde{\Delta} = \tilde{\Delta}(\varepsilon)$ *such that* $\tilde{\Delta} \to 0+$ *as* $\varepsilon \to 0$ *and such that*

$$y(t;\varepsilon) = 1 + \varepsilon^{1/3} f_1(\tilde{t}) + o(\varepsilon^{1/3}) \tag{83}$$

uniformly for $\tilde{t} \in [\tilde{t}_0 - \tilde{\nu},\ \tilde{t}_0 - \tilde{\Delta}]$. *Furthermore, without loss of generality, we can assume*

$$\frac{\varepsilon^{1/3}|\log \varepsilon|}{\tilde{\Delta}^2} \to 0 \quad \text{as } \varepsilon \to 0. \tag{84}$$

PROPOSITION 14.

$$\frac{dr}{dt^*}(0) = o(\varepsilon^{1/3}). \tag{85}$$

Proof. Tentatively assume γ is a positive constant and that there are arbitrarily small ε-values for which

$$\frac{dr}{dt^*} < -\gamma\varepsilon^{1/3}; \tag{86}$$

throughout the proof assume ε is chosen from this set. It is easily shown that for ε sufficiently small,

$$r'(-1) < \frac{-\gamma\varepsilon^{1/3}}{4}, \tag{87}$$

$$r(-1) > \frac{\gamma\varepsilon^{1/3}}{2}. \tag{88}$$

If ε is sufficiently small, $-\varepsilon^{2/3} > -\gamma\varepsilon^{1/3}/4$, and for such ε denote by t_1^* the first t^*-value to the left of -1, where dr/dt^* has increased to $-\varepsilon^{2/3}$. Kevorkian and Cole establish asymptotic properties for $g_0(t^*)$ which imply the existence of constants C_1, C_2, C_3 such that if $t^* \leqq -1$,

$$g_0(t^*) \geqq C_1, \tag{89}$$

$$g_0'(t^*) < \frac{-C_2}{t^{*2}}, \tag{90}$$

$$1 - (g_0(t^*))^2 < \frac{2C_3}{(-t^*)}. \tag{91}$$

These and (a) of Proposition 12 used in (11) give

$$\frac{d^2r}{dt^{*2}} > \frac{2C_1C_2}{t^{*2}} + \left(\frac{2C_3}{(-t^*)}\right) \cdot \left(\frac{dr}{dt^*}\right) - 3\varepsilon, \qquad t^* \in [t_1^*, -1], \tag{92}$$

and at t_1^* it is necessary that the left side of (92) be nonpositive. Hence, after multiplying the right side of (92) by $(t_1^*)^2$, there results a quadratic in t_1^*, and t_1^* must be at least as negative as the negative root of the quadratic. Assuming γ is small enough to ensure

$$\frac{3\gamma C_1C_2}{C_3^2} < 1, \tag{93}$$

elementary computations yield

$$t_1^* < \left(\frac{-3C_1C_2\gamma}{8C_3}\right)\varepsilon^{-1/3}. \tag{94}$$

Relating t^*-values and $\tilde{t}$-values with (5) and (6), it is readily seen that $\tilde{t}_0-\tilde{\Delta}$ lies to the right of t_1^* if ε is sufficiently small, and hence at the t^*-value corresponding to $\tilde{t}_0-\tilde{\Delta}$, dr/dt^* is more negative than $-\varepsilon^{2/3}$ and r is at least as positive as $\gamma\varepsilon^{1/3}/2$. Now use the asymptotic behavior of $g_0(t^*)$ as recorded by Kevorkian and Cole [4, form. (2.6.24)]—see (25), and evaluate g_0 at the t^* corresponding to $\tilde{t}_0-\tilde{\Delta}$:

$$g_0\left(-\varepsilon^{-1/3}\tilde{\Delta}\left[1+\frac{\varepsilon^{1/3}|\log\varepsilon|}{9\tilde{\Delta}}\right]\right)=1+\left(\frac{\varepsilon^{1/3}}{-\tilde{\Delta}}\right)\left\{1+O\left(\frac{\varepsilon^{1/3}|\log\varepsilon|}{9\tilde{\Delta}}\right)\right\}+O\left\{\frac{\log|-\varepsilon^{1/3}\tilde{\Delta}(1+o(1))|}{[\varepsilon^{-1/3}\tilde{\Delta}(1+o(1)]^2}\right\}. \tag{95}$$

With (84), this leads to

$$\begin{aligned} y(\tilde{t}=\tilde{t}_0-\tilde{\Delta}) &= 1-\frac{\varepsilon^{1/3}}{\tilde{\Delta}}+o(\varepsilon^{1/3})+r(\tilde{t}=\tilde{t}_0-\tilde{\Delta}) \\ &> 1-\frac{\varepsilon^{1/3}}{\tilde{\Delta}}+\frac{\gamma\varepsilon^{1/3}}{2}+o(\varepsilon^{1/3}). \end{aligned} \tag{96}$$

On the other hand, from Proposition 13 and (30),

$$\begin{aligned} y(\tilde{t}=\tilde{t}_0-\tilde{\Delta}) &= 1+\varepsilon^{1/3}f_1(\tilde{t}_0-\tilde{\Delta})+o(\varepsilon^{1/3}) \\ &= 1-\frac{\varepsilon^{1/3}}{\tilde{\Delta}}+o(\varepsilon^{1/3}). \end{aligned} \tag{97}$$

Comparing (96) and (97) yields the needed contradiction and the proposition is proved.

PROPOSITION 15. *Let $t_{\tilde{\Delta}}^*$ denote the t^*-value corresponding to $\tilde{t}=\tilde{t}_0-\tilde{\Delta}$. Then*

$$r(t^*)=o(\varepsilon^{1/3}) \tag{98}$$

uniformly for $t^*\in[t_{\tilde{\Delta}}^*, 0]$.

Proof. The proof is similar to that of Proposition 14. Thus, let γ denote a positive constant and assume tentatively the existence of $t_1^*\in[t_{\tilde{\Delta}}^*, 0]$, where $r(t_1^*)\geqq\gamma\varepsilon^{1/3}$. Then the Mean Value Theorem produces a $t_\alpha^*\in(t_{\tilde{\Delta}}^*, 0)$, where $r'(t_\alpha^*)$ is at least as negative as $(\gamma\varepsilon^{1/3})/(2t_{\tilde{\Delta}}^*)$, and hence (recall $t_{\tilde{\Delta}}^*=-\varepsilon^{-1/3}\tilde{\Delta}(1+o(1))$) for all sufficiently small ε,

$$r'(t_\alpha^*)<-\varepsilon^{2/3}, \tag{99}$$

and of course

$$r'(t_\alpha^*)>\gamma\varepsilon^{1/3}/2. \tag{100}$$

The remainder of the proof is essentially like that of Proposition 14.

COROLLARY. *Let $\eta^*=o(\varepsilon^{-1/3})$. Then*

$$y=g_0(t^*)+o(\varepsilon^{1/3}) \tag{101}$$

uniformly for $t^*\in[-\eta^*, 0]$.

Proof. In the proof of Proposition 15, replace $\tilde{\Delta}$ by $\max\{\varepsilon^{1/3}\eta^*, \tilde{\Delta}\}$.

We now turn our attention to the behavior of $r(t^*)$ to the right of $t^*=0$. The goal is to sharpen the estimate of Proposition 6 in [8].

PROPOSITION 16. *Let $1\ll\eta\ll\varepsilon^{-2/3}$. Then both $r(t^*)$ and $r'(t^*)$ are $o(\varepsilon^{1/3})$ uniformly for $t^*\in[0, \eta]$.*

Proof. As in [8], begin with the identities

$$\frac{dr}{dt^*}-(1-g_0^2)r=\frac{dr}{dt^*}(0)+\int_0^{t^*}\left\{-\left(\frac{dg_0}{d\nu}\right)r^2-\varepsilon y-[2g_0+r]r\left(\frac{dr}{d\nu}\right)\right\}d\nu, \tag{102}$$

$$\begin{aligned} r(t^*)=&\left(\frac{dr}{dt^*}(0)\right)\left\{\exp\int_0^{t^*}(1-g_0^2)\,ds\right\}\int_0^{t^*}\left\{\exp\left[-\int_0^{\xi}(1-g_0^2)\,ds\right]\right\}d\xi \\ &+\left\{\exp\int_0^{t^*}(1-g_0^2)\,ds\right\}\int_0^{t^*}\left\{\exp\left[-\int_0^{\xi}(1-g_0^2)\,ds\right]\right\} \\ &\cdot\left\{\int_0^{\xi}\left[-\left(\frac{dg_0}{d\nu}\right)r^2-\varepsilon y(2g_0+r)r\left(\frac{dr}{d\nu}\right)\right]d\nu\right\}\Bigg\}d\xi \end{aligned} \tag{103}$$

At $t^*=0$, both r and r' are $o(\varepsilon^{1/3})$, and this estimate certainly cannot be destroyed on any finite interval due to the presence of products of r and r' in the integrands. Choose the finite interval to be $[0, 4]$. So,

$$r(t^*)=o(\varepsilon^{1/3}), \tag{104}$$

$$r'(t^*)=o(\varepsilon^{1/3}), \tag{105}$$

uniformly on $[0, 4]$.

To continue these estimates to the right of $t^*=4$, multiply (11) by dr/dt^*:

$$\begin{aligned} \frac{d}{dt^*}\left(\frac{dr}{dt^*}\right)^2=&-3\left(\frac{dr}{dt^*}\right)^2+\left(\frac{dr}{dt^*}\right) \\ &\cdot\left[-(2g_0')\cdot(2g_0+r)r+(-2(y^2-1)+3)\left(\frac{dr}{dt^*}\right)-2\varepsilon y\right]. \end{aligned} \tag{106}$$

If the expression in the square braces were absent from (106), then $|dr/dt^*|$ would decay exponentially fast and the $o(\varepsilon^{1/3})$ estimate for $r(4)$, $r'(4)$ would persist for all $t^*>4$. The task is thus to control the term in the square braces, at least until t^* reaches η. To this end, let k denote any positive constant and choose ε small enough to ensure neither $|r(4)|$ nor $|r'(4)|$ exceeds $k\varepsilon^{1/3}/2$. Then, for as long as

$$|r(t^*)|\leqq 2k\varepsilon^{1/3}, \tag{107}$$

$$|r'(t^*)|\leqq k\varepsilon^{1/3} \tag{108}$$

hold to the right of $t^*=4$, simple computations yield

$$\left(\frac{dr}{dt^*}\right)^2<(k\varepsilon^{1/3})^2. \tag{109}$$

(To arrive at (109), use (a) and (d) of Proposition 12, together with the facts that g_0 is within .001 of -2 and $g_0'(t^*)$ has magnitude smaller than .001 for $t^*\geqq 4$.) Thus the weak inequalities (107), (108) imply the strict inequality (109), so it is now only necessary to exclude the possibility that equality in (107) occurs at some t^*-value in $[4, \eta]$. Assume tentatively that $r(t^*)$ reaches $2k\varepsilon^{1/3}$ for the first time at $t^*=t_2^*$, and denote by t_1^* the last t^*-value previous to t_2^* that $r=0$, or $t_1^*=4$, whichever is greater. (The case $r(t_2^*)=-2k\varepsilon^{1/3}$ is considered below.)

Observe that on any subinterval of $[t_1^*, t_2^*]$ where $r'(t^*)$ happens to be as large as $5\varepsilon/3$, it is easily shown that if ε is sufficiently small, the expression in the braces in

(106) is negative, and from this it follows immediately that if dr/dt^* ever exceeds $5\varepsilon/3$ on $[t_1^*, t_2^*]$ it must do so at t_1^*. Now integrate the inequality, obtained from (106),

$$\frac{d}{dt^*}\left(\frac{dr}{dt^*}\right)^2 < -3\left(\frac{dt}{dt^*}\right)^2 \tag{110}$$

for as long as $dr/dt^* \geqq 5\varepsilon/3$ on $[t_1^*, t_2^*]$. The trivial computation yields

$$\begin{aligned} r(t^*) &\leqq r(t_1^*) + r'(t_1^*)\frac{2}{3}\left[1 - \exp\left(\frac{-3(t^* - t_1^*)}{2}\right)\right] \\ &< \frac{3k\varepsilon^{1/3}}{2}. \end{aligned} \tag{111}$$

Obviously, once dr/dt^* has dropped to $5\varepsilon/3$, it can never again equal $5\varepsilon/3$, and so any further increase in r on the interval $[t_1^*, t_2^*]$ is limited to $5\varepsilon/3 \cdot (t_2^* - t_1^*)$, and this cannot exceed $5\varepsilon/3 \cdot \eta$. Consequently,

$$\begin{aligned} r(t_2^*) &\leqq \frac{3k\varepsilon^{1/3}}{2} + \frac{5\varepsilon}{3}\eta \\ &< 2k\varepsilon^{1/3} \end{aligned} \tag{112}$$

if ε is sufficiently small. This contradicts the definition of t_2^*.

It is even easier to prove that $r(t^*) > -2k\varepsilon^{1/3}$ for $t^* \in [4, \eta]$ since if r, r' are both nonpositive, all the terms in the braces of (106) are nonpositive.

It has now been shown that for arbitrary positive k, (107), (108) are valid for all $t^* \in [4, \eta]$, provided ε is sufficiently small. Combining this with the result on the interval $[0, 4]$ completes the proof.

From the corollary to Proposition 15 and from Proposition 16 we have the following theorem.

THEOREM B. *Let* $0 \ll \eta^* \ll \varepsilon^{-1/3}$ *and* $0 \ll \eta \ll \varepsilon^{-2/3}$. *Then*

$$y = g_0(t^*) + o(\varepsilon^{1/3}) \tag{113}$$

uniformly for $-\eta^* \leqq t^* \leqq \eta$.

5. Outer expansion analysis. In this section we prove that the leading term of the outer asymptotic expansion, $u_0(t)$, gives an $O(\varepsilon^{1/3})$ approximation to y on an explicit domain of uniform validity.

PROPOSITION 17. *Let* $\varepsilon^{2/3} \ll \eta_2 \leqq \varepsilon^{1/3}$. *Then*

$$y(\delta - \eta_2) = u_0(\delta - \eta_2) + o(\varepsilon^{1/3}), \tag{114}$$

$$h(t) \leqq o(\varepsilon^{1/3}) \quad \textit{uniformly on} \left[\frac{-T}{2} + \delta, \delta - \eta_2\right]. \tag{115}$$

Proof. If ε is sufficiently small, $\delta - \eta_2$ is in the domain of uniform validity of the transition expansion; see Theorem A, (4), and (6). An elementary computation involving the asymptotic behavior of f_1 (see (32)) gives

$$y(\delta - \eta_2) = 1 + \varepsilon^{1/3}(\varepsilon^{-2/3}\eta_2 - \tilde{t}_0)^{1/2} + o(\varepsilon^{1/3}). \tag{116}$$

On the other hand, the asymptotic behavior of u_0 described by Kevorkian and Cole's equation [4, eq. (2.6.15)] yields

$$y(\delta - \eta_2) = 1 + \varepsilon^{1/3}(\varepsilon^{-2/3}\eta_2 - \tilde{t}_0)^{1/2} + o(\varepsilon^{1/3}) + h(-\eta_2 + \delta). \tag{117}$$

Combining (116) and (117) yields (114).

To show h is bounded above by $o(\varepsilon^{1/3})$ uniformly on $[-T/2+\delta, -\eta_2+\delta]$, note $y(-T/2+\delta)=0$ and $u_0(-T/2+\delta)>0$, so $h(T/2+\delta)$ is negative. It is now easy to show that the assumptions that $h(\delta-\eta_2)$ is nonnegative and that h exceeds $h(\delta-\eta_2)$ anywhere on the interval lead to a contradiction. If $h(\delta-\eta_2)$ is negative, a similar argument shows $h(t)$ is negative on $[-T/2+\delta, \delta-\eta_2]$. This completes the proof.

The next two propositions give a proof of a classical result. The method of proof in Proposition 19 is used in Proposition 20.

PROPOSITION 18.

$$T \leqq (3-2\log 2)+o(\varepsilon^{1/3}). \tag{118}$$

Proof. Only those small ε-values for which $-T/2 > -\frac{3}{2}+\log 2$ need be considered. Let Δt_ε satisfy

$$|\log \varepsilon| \leqq \frac{\Delta t_\varepsilon}{\varepsilon} \ll \varepsilon^{-2/3}. \tag{119}$$

From Theorem B and the antiperiodic behavior of the oscillations, and equation (2.6.72) of [4], there follows

$$\begin{aligned} y\left(\frac{-T}{2}+\delta+\Delta t_\varepsilon\right) &= -g_0\left(\frac{\Delta t_\varepsilon}{\varepsilon}\right)+o(\varepsilon^{1/3}) \\ &= 2+o(\varepsilon^{1/3}), \end{aligned} \tag{120}$$

and thus

$$h\left(\frac{-T}{2}+\delta+\Delta t_\varepsilon\right) = 2+o(\varepsilon^{1/3})-u_0\left(\frac{-T}{2}+\delta+\Delta t_\varepsilon\right). \tag{121}$$

Noting that u_0' is increasing on the interval between $-\frac{3}{2}+\log 2$ and $-T/2+\delta+\Delta t_\varepsilon$, that $u_0'(-\frac{3}{2}+\log 2)=-\frac{2}{3}$, and that $u_0(-\frac{3}{2}+\log 2)=2$, we can easily show with the Mean Value Theorem that

$$h\left(\frac{-T}{2}+\delta+\Delta t_\varepsilon\right) > o(\varepsilon^{1/3})+\frac{2}{3}\left[\left(\frac{-T}{2}+\delta+\Delta t_\varepsilon\right)-\left(-\frac{3}{2}+\log 2\right)\right]. \tag{122}$$

Now use (115) to complete the proof.

PROPOSITION 19.

$$T \leqq (3-2\log 2)+o(\varepsilon^{1/3}). \tag{123}$$

Proof. Assume tentatively that there is a positive constant γ and a set $E(\gamma)$ of ε-values which contains arbitrarily small values of ε for which

$$\frac{-T}{2} < -\frac{3}{2}+\log 2-\frac{\gamma\varepsilon^{1/3}}{2}. \tag{124}$$

Let

$$t_\varepsilon = \frac{-T}{2}+\delta+\Delta t_\varepsilon. \tag{125}$$

Then from (121),

$$h(t_\varepsilon) = 2+o(\varepsilon^{1/3})-u_0(t_\varepsilon). \tag{126}$$

If ε is sufficiently small,

$$t_\varepsilon < -\frac{3}{2}+\log 2-\frac{7\gamma\varepsilon^{1/3}}{16} \tag{127}$$

and since $u_0(-\frac{3}{2}+\log 2)=2$, (127) and (126) imply the existence of a positive constant k that is *independent of* ε *for all sufficiently small* ε, such that

$$h(t_\varepsilon)\leqq -k\varepsilon^{1/3}; \tag{128}$$

that is, h is negative and bounded away from zero by $O(\varepsilon^{1/3})$. The method of proof proceeds by showing that h continues to be negative and bounded away from zero to the right of t_ε, even penetrating the region of uniform validity of the transition expansion. But estimates from the latter prevent h from being negative and bounded away from zero by $O(\varepsilon^{1/3})$, thus proving the proposition.

To prove that (128) implies that h continues to be negative and bounded away from zero by $O(\varepsilon^{1/3})$, it is necessary to use technical details that are straightforward computations, and only a proof outline will be given. Begin with the differential equation for h written in the form

$$\begin{aligned}\frac{d^2h}{dt^2}=&-\left(\frac{1}{\varepsilon}\right)(y^2-1)\frac{dh}{dt}+\left(\frac{1}{2\varepsilon}\right)\frac{(1+u_0y)h}{(u_0^2-1)}\\&+\left(\frac{1}{2\varepsilon}\right)\frac{(1+u_0y)h}{(u_0^2-1)}\left\{1+\frac{u_0(1+u_0^2)2\varepsilon}{(u_0^2-1)^2h}\right\}\end{aligned} \tag{129}$$

and note first that Proposition 16, together with the exponential decay of g_0', implies $h'(t_\varepsilon)$ is bounded above by $o(\varepsilon^{-2/3})$. Thus, if $h'(t_\varepsilon)>\varepsilon^{1/3}$, we use the first term on the right side of (129) (which is approximately $-3h'/\varepsilon$) to show $h'(t)$ drops to $\varepsilon^{1/3}$ so fast that h is still negative and bounded away from zero by $O(\varepsilon^{1/3})$. Further, the t-value where h' reaches $\varepsilon^{1/3}$ is, like t_ε, on the left of and bounded away from $-\frac{3}{2}+\log 2$ by $O(\varepsilon^{1/3})$.

Now continue, and with the help of the second term on the right side of (129), show that h' drops to zero and becomes negative in a time interval of $O(\varepsilon)$. Thus, if h' was not negative at t_ε, it becomes so at a t-value that is, like t_ε itself, on the left of and bounded away from $-\frac{3}{2}+\log 2$ by $O(\varepsilon^{1/3})$. And clearly h is still negative and bounded away from zero by $O(\varepsilon^{1/3})$. Obviously, $h'(t)$ remains negative as long as the term in the braces of (129) remains positive, so the last step is to follow the sign of the term in the braces.

Note that $u_0(t)$ decreases to $u_0(0)=1$, as t increases. This leads to the conclusion that the term in the braces remains positive as long as

$$\frac{5\varepsilon}{(u_0-1)^2|h|}\leqq 1. \tag{130}$$

It turns out that the estimates above lead to the conclusion

$$h(t)\leqq\frac{-k\varepsilon^{1/3}}{4} \tag{131}$$

at least until (130) fails. It has been shown [9, eq. 50] that $t>-(u_0-1)^2$ when $-\frac{3}{2}+\log 2\leqq t<0$. It follows that (130) remains valid at least until

$$t=\frac{-20\varepsilon^{2/3}}{k}, \tag{132}$$

if ε is sufficiently small, and so the tentative assumption implies (131) holds for

$$t\in\left[t_\varepsilon,\frac{-20\varepsilon^{2/3}}{k}\right]. \tag{133}$$

However, consider the t-value corresponding to $\tilde{t} = \tilde{t}_0 - \tilde{\nu}$, where $\tilde{\nu}$ is defined in Theorem A. That is, $t = \varepsilon^{2/3}(\tilde{t}_0 - \tilde{\nu}) - \varepsilon|\log \varepsilon|/6$. It is easily verified from (79) that

$$y(\tilde{t} = \tilde{t}_0 - \tilde{\nu}) = 1 + \varepsilon^{1/3}(-\tilde{t}_0 + \tilde{\nu})^{1/2} + o(\varepsilon^{1/3}), \tag{134}$$

and from (6), (8), and the small-t behavior of u_0 that

$$y\left(t = \varepsilon^{2/3}(\tilde{t}_0 - \tilde{\nu}) - \frac{\varepsilon|\log \varepsilon|}{6}\right) = 1 + \varepsilon^{1/3}(-\tilde{t}_0 + \tilde{\nu})^{1/2} + o(\varepsilon^{1/3}) + h\left(\tilde{t}_0 - \tilde{\nu} - \frac{\varepsilon|\log \varepsilon|}{6}\right). \tag{135}$$

Thus

$$h\left(\varepsilon^{2/3}(\tilde{t}_0 - \tilde{\nu}) - \frac{\varepsilon|\log \varepsilon|}{6}\right) = o(\varepsilon^{1/3}),$$

contradicting (131), since $t = \varepsilon^{2/3}(\tilde{t}_0 - \tilde{\nu}) - \varepsilon|\log \varepsilon|/6$ is obviously in the interval (133) when ε is sufficiently small.

THEOREM C. *Let $\eta_1(\varepsilon)$, $\eta_2(\varepsilon)$ satisfy*

$$\varepsilon|\log \varepsilon| \leqq \eta_1 \ll \varepsilon^{1/3} \quad \textit{and} \quad \varepsilon^{2/3} \ll \eta_2 \ll \varepsilon^{1/3}. \tag{136}$$

Then

$$y(t; \varepsilon) = u_0(t) + o(\varepsilon^{1/3}) \tag{137}$$

uniformly for

$$t \in \left[\frac{-T}{2} + \delta + \eta_1, \delta - \eta_2\right]. \tag{138}$$

Proof. From Proposition 16, $\delta + \eta_1$ is in the domain where $r(t^*) = o(\varepsilon^{1/3})$ and where $g_0(t^*) = 2 + o(\varepsilon^{1/3})$. Using the antiperiodic behavior of y gives

$$y\left(\frac{-T}{2} + \delta + \eta_1\right) = 2 + o(\varepsilon^{1/3}). \tag{139}$$

Using $T = 3 - 2\log 2 + o(\varepsilon^{1/3})$ gives

$$y\left(\frac{-T}{2} + \delta + \eta_1\right) = 2 + o(\varepsilon^{1/3}) + h\left(\frac{-T}{2} + \delta + \eta_1\right), \tag{140}$$

and so

$$h\left(\frac{-T}{2} + \delta + \eta_1\right) = o(\varepsilon^{1/3}).$$

Next let k denote any positive constant and assume tentatively that h falls below $-k\varepsilon^{1/3}$ on the interval (138). Because of (139) and (114) there must be a t-value between $-T/2 + \delta + \eta_1$ and $\delta - \eta_2$, where h is less than $-k\varepsilon^{1/3}$ and where h' vanishes. The continuation argument of Proposition 19 is now applied and the same contradiction reached; the theorem is proved.

6. Summary. There remains the pleasant and easy task of collecting together the results of Theorems A–C, to confirm that together the explicit domains of validity of

the terms of the inner, outer, and transition expansions studied above indeed overlap in such a way as to cover the half period $t \in [-T/2+\delta, \delta]$. All domains are expressed in terms of the outer variable t, and are shown schematically in Fig. 1.

THEOREM D. *Let $\tilde{\Delta}$ be as defined in Proposition* 13, *and let η, η_1, η_2, $\tilde{\nu}$, and η^* satisfy*

(a) $\varepsilon|\log \varepsilon| \ll \eta_1 \ll \varepsilon\eta \ll \varepsilon^{1/3}$,

(b) $\varepsilon^{2/3} \ll \eta_2 \ll \varepsilon^{2/3}\tilde{\nu} + \varepsilon|\log \varepsilon|/9 \ll \varepsilon^{1/3}$,

(c) $\sqrt{\varepsilon^{5/3}|\log \varepsilon|} \ll \varepsilon^{2/3}\tilde{\Delta} + \varepsilon|\log \varepsilon|/9 \ll \varepsilon\eta^* \ll \varepsilon^{2/3}$.

Then, uniformly and within an $o(\varepsilon^{1/3})$ error, y is approximated by the following expressions:

$$-g_0\left(t^* + \frac{T}{2\varepsilon}\right) \quad \text{for } t \in \left[\frac{-T}{2} + \delta, \frac{-T}{2} + \delta + \varepsilon\eta\right], \tag{141}$$

$$u_0(t) \quad \text{for } t \in \left[\frac{-T}{2} + \delta + \eta_1, \delta - \eta_2\right], \tag{142}$$

$$1 + \varepsilon^{1/3} f_1(\tilde{t}) \quad \text{for } t \in \left[\delta - \frac{\varepsilon|\log \varepsilon|}{9} - \varepsilon^{2/3}\tilde{\nu}, \delta - \frac{\varepsilon|\log \varepsilon|}{9} - \varepsilon^{2/3}\tilde{\Delta}\right], \tag{143}$$

$$g_0(t^*) \quad \text{for } t \in [\delta - \varepsilon\eta^*, \delta]. \tag{144}$$

Proof. If η satisfies (a) and if η^* satisfies (c), then (141) and (144) follow immediately from Theorem B; of course, a translation, based on the antiperiodicity of y, is needed for (141). If η_1, η_2 satisfy (a) and (b), then (142) follows from Theorem C. Finally, if $\tilde{\nu}$ satisfies (b) then (143) follows from Theorem A. Obviously the intervals in (141) through (144) overlap and their union is $[-T/2+\delta, \delta]$.

REFERENCES

[1] M. CARTWRIGHT, *Van der Pol's equation for relaxation oscillations*, in Contributions to the Theory of Nonlinear Oscillations, Ann. of Math. Stud., 29, Princeton University Press, Princeton, NJ, 1952, pp. 3-18.

[2] A. A. DORODNYCIN, *Asymptotic solution of Van der Pol's equation*, Prikl. Mat. Mekh., 11 (1947), pp. 313-328; Trans. Amer. Math. Soc. 88 (1953), pp. 1-24.

[3] J. HAAG, *Etude asymptotic des oscillations de relaxation*, Ann. Sci. Ecole Norm. Sup. (3), 60 (1944), pp. 73-117.

[4] J. KEVORKIAN AND J. D. COLE, *Perturbation Methods in Applied Mathematics*, Springer-Verlag, Berlin, New York, 1981.

[5] P. A. LAGERSTROM AND R. G. CASTEN, *Basic concepts underlying singular perturbation techniques*, SIAM Rev., 14 (1972), pp. 63-120.

[6] J. S. LANGER, *Existence of needle crystals in local models of solidification*, Phys. Rev. A, 33 (1986), pp. 435-441.

[7] J. E. LITTLEWOOD, *Littlewood's Miscellany*, B. Bollobas, ed., Cambridge University Press, Cambridge, 1986.

[8] A. D. MACGILLIVRAY, *On the leading term of the inner asymptotic expansion of Van der Pol's equation*, SIAM J. Appl. Math., 43 (1983), pp. 594-612.

[9] ———, *On the leading term of the outer asymptotic expansion of Van der Pol's equation*, SIAM J. Appl. Math., 43 (1983), pp. 1221-1239.

[10] ———, *On the transition asymptotic expansion of Van der Pol's equation*, preprint, 1987.

[11] ———, *Matching of the inner, outer, and transition expansions of Van der Pol's equation*, preprint, 1987.

[12] E. F. MISHCHENKO AND B. KH. ROZOV, *Differential equations with small parameters and relaxation oscillations*, Plenum Press, New York, 1980.

[13] R. E. O'MALLEY, *Introduction to Singular Perturbations*, Academic Press, New York, 1974.

[14] P. G. SAFFMAN, *Viscous fingering in Hele-Shaw cells*, J. Fluid Mech., 173 (1986), pp. 73-94.

[15] J. J. STOKER, *Nonlinear vibrations in mechanical and electrical systems*, Interscience, New York, 1950.

[16] D. W. STORTI AND R. H. RAND, *Dynamics of two strongly coupled relaxation oscillators*, SIAM J. Appl. Math., 46 (1986), pp. 56-67.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 241-261, January 1990
© 1990 Society for Industrial and Applied Mathematics
013

UNIFORM ASYMPTOTIC EXPANSIONS OF A CLASS OF INTEGRALS IN TERMS OF MODIFIED BESSEL FUNCTIONS, WITH APPLICATION TO CONFLUENT HYPERGEOMETRIC FUNCTIONS*

N. M. TEMME†

Abstract. The integral

$$F_\lambda(z,\alpha)=\int_0^\infty t^{\lambda-1}\,e^{-zt-\alpha/t}f(t)\,dt$$

is considered for large values of the real parameter z; α and λ are uniformity parameters in $[0,\infty)$. The asymptotic expansion is given in terms of the modified Bessel function $K_\lambda(2\sqrt{\alpha z})$. The asymptotic nature of the expansion is discussed and error bounds are constructed for the remainders in the expansions. An example is given for confluent hypergeometric or Whittaker functions. In this example the integrals are transformed to standard forms and the mappings are investigated.

Key words. uniform asymptotic expansions of integrals, modified Bessel function, confluent hypergeometric function, Whittaker function, construction of error bounds, transformation to standard form

AMS(MOS) subject classifications. 41A60, 30E15, 33A20

1. Introduction.

We consider integrals of the form

$$F_\lambda(z,\alpha)=\int_0^\infty t^{\lambda-1}\,e^{-zt-\alpha/t}f(t)\,dt, \tag{1.1}$$

which reduces to a modified Bessel function in the case that f is a constant. We have

$$2(\alpha/z)^{\lambda/2}K_\lambda(2\sqrt{\alpha z})=\int_0^\infty t^{\lambda-1}\,e^{-zt-\alpha/t}\,dt. \tag{1.2}$$

The integral in (1.1) is considered with $\alpha,\ \lambda\geqq 0$ and large positive values of z. We aim to derive asymptotic expansions for $F_\lambda(z,\alpha)$ that hold uniformly with respect to both α and λ in the interval $[0,\infty)$. To handle the transition of the case $\alpha=0$ to $\alpha>0$, the modified Bessel function (1.2) is needed. Observe that when $\alpha=0$ the essential singularity in the integrand of (1.1) disappears and that (1.1) becomes a more familiar Laplace integral, which can be expanded by using Watson's lemma.

First we consider fixed values of λ. To describe the asymptotic features we introduce the positive number β defined by

$$\beta=\sqrt{\alpha/z}. \tag{1.3}$$

The saddle points of $\exp(-zt+\alpha/t)$ are located at $t=\pm\beta$. When β is bounded away from zero, we can use the familiar Laplace method, since at the point $t=\beta$ the integrand has the form of a Gaussian function. When, however, $\alpha\to 0$, that is, $\beta\to 0$, the internal saddle point coalesces with the point $t=0$, where the argument of the exponential function has a pole. In addition, there is an algebraic singularity (if $\lambda\neq 1$), but the influence of the essential singularity due to the pole is more significant. Observe that in the limit $\alpha=0$, as mentioned earlier, the pole disappears; also, both saddle points coalesce with the pole. These asymptotic features are typical for certain integrals defining Bessel functions. For this reason the modified Bessel function in (1.2) serves

* Received by the editors May 5, 1988; accepted for publication (in revised form) February 27, 1989.
† Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, the Netherlands.

as a basic approximant in the uniform asymptotic expansions in this paper. In § 4 we show how an integral with the same phenomena can be transformed into the standard form (1.1).

The integral in (1.1) is the simplest case with the asymptotic features described above, especially when the parameters are in the indicated intervals. We apply the results to a confluent hypergeometric function. By allowing different intervals of integration, say a contour in the complex plane, we can also consider negative values of α. Then the ordinary Bessel function $J_\nu(z)$ shows up. This case is more difficult, but the applications are very interesting in the theory of special functions.

Consider as an analogue of (1.1) a loop integral in the form

$$G_\lambda(z,\alpha)=\frac{1}{2\pi i}\int_{-\infty}^{(0^+)} t^{-\lambda-1}\,e^{zt+\alpha/t}f(t)\,dt. \tag{1.4}$$

This notation means that the contour of integration starts from $-\infty$, arg $t=-\pi$, describes a circle counterclockwise around the origin, and returns to $-\infty$, arg $t=+\pi$. The integral (1.4) has the modified Bessel function $I_\lambda(2\sqrt{\alpha z})$ as approximant. When $f=1$ we have

$$G_\lambda(z,\alpha)=(z/\alpha)^{\lambda/2}I_\lambda(2\sqrt{\alpha z}). \tag{1.5}$$

When α is negative this function is an ordinary J-Bessel function. In [2] and [6] integrals of the type (1.4) are treated and the method is used for obtaining a uniform expansion of Laguerre polynomials. We plan to return to this problem in a future paper.

The starting point (1.1) is of interest since it has a real interval of integration. Thus the transformation to the standard form (1.1) involves a real mapping. This makes the first steps of the analysis rather simple, since we do not need to trace the transformed contour in the complex plane. For studying the asymptotic nature of the expansion, we use complex variables, however.

The plan of the paper is as follows. In § 2 we construct a series expansion based on an integration by parts procedure, and we give estimates for the remainder in the expansion. In § 3 we consider an expansion that is based on expanding f at the internal saddle point. In § 4 we give an application to confluent hypergeometric functions. In § 5 the parameter λ is considered as a second uniformity parameter in $[0,\infty)$, and again we apply the methods on a confluent hypergeometric function. Especially, we pay attention to the mappings needed for a transformation to the standard form.

Terminology. We call a parameter *fixed* when it does not depend on the parameters z, α, λ. $\Re z=x$, $\Im z=y$ are the real and imaginary part of $z=x+iy$.

2. An integration by parts procedure. The procedure of this section takes into account both saddle points $\pm\beta$ of the exponential function (where β is given in (1.3)), although $-\beta$ lies outside the interval of integration. For this reason we assume that f is also defined at negative values of its argument, and that f is sufficiently smooth for the operations to be used here. Further conditions on f will be given later.

2.1. Construction of the formal series. The first step is the representation

$$f(t)=a_0+b_0(t-\beta)+(t-\beta^2/t)g(t), \tag{2.1}$$

where a_0, b_0 follow from substitution of $t=\pm\beta$. We have

$$a_0=f(\beta),\qquad b_0=\frac{1}{2\beta}[f(\beta)-f(-\beta)].$$

Inserting (2.1) into (1.1) we obtain

$$F_\lambda(z,\alpha)=a_0A_\lambda(z,\beta)+b_0B_\lambda(z,\beta)+F_\lambda^{(1)}(z,\alpha),$$

where A_λ, B_λ are combinations of the modified Bessel functions introduced in (1.2). It is straightforward to verify that

$$A_\lambda(z,\beta)=2\beta^\lambda K_\lambda(2\beta z), \quad B_\lambda(z,\beta)=2\beta^{\lambda+1}[K_{\lambda+1}(2\beta z)-K_\lambda(2\beta z)]. \tag{2.2}$$

An integration by parts gives

$$\begin{aligned} F_\lambda^{(1)}(z,\alpha) &= -\frac{1}{z}\int_0^\infty t^\lambda g(t)\, d\exp(-z(t+\beta^2/t)) \\ &= \frac{1}{z}\int_0^\infty t^{\lambda-1}\exp(-z(t+\beta^2/t))f_1(t)\,dt \end{aligned}$$

with

$$f_1(t)=t^{1-\lambda}\frac{d}{dt}[t^\lambda g(t)]=\lambda g(t)+tg'(t).$$

We see that $zF_\lambda^{(1)}(z,\alpha)$ is of the same form as $F_\lambda(z,\alpha)$. The above procedure can now be applied to $zF_\lambda^{(1)}(z,\alpha)$, and we obtain for (1.1) the formal expansion

$$F_\lambda(z,\alpha)\sim A_\lambda(z,\beta)\sum_{s=0}^\infty a_s z^{-s}+B_\lambda(z,\beta)\sum_{s=0}^\infty b_s z^{-s} \quad \text{as } z\to\infty, \tag{2.3}$$

where we define inductively $f_0=f$, $g_0=g$ and for $s=1,2,\cdots$

$$\begin{aligned} &f_s(t)=t^{1-\lambda}\frac{d}{dt}[t^\lambda g_{s-1}(t)]=a_s+b_s(t-\beta)+\left(t-\frac{\beta^2}{t}\right)g_s(t), \\ &a_s=f_s(\beta), \qquad b_s=\frac{1}{2\beta}[f_s(\beta)-f_s(-\beta)]. \end{aligned} \tag{2.4}$$

Remark 2.1. As mentioned earlier, for this procedure we need function values of f and derivatives at negative values, although the integral (1.1) is defined only for t-values in $[0,\infty)$. When we consider analytic functions f, as we do later, we assume that f is analytic in a domain Ω in the complex plane that contains the real line. When, however, f is supposed to belong to $C^k[0,\infty)$, we assume in the above procedure that f has been smoothly continued on $(-\infty,0]$.

2.2. The remainder of the expansion. We introduce a remainder for the expansion in (2.3) by writing

$$F_\lambda(z,\alpha)=A_\lambda(z,\beta)\sum_{s=0}^{n-1} a_s z^{-s}+B_\lambda(z,\beta)\sum_{s=0}^{n-1} b_s z^{-s}+z^{-n}R_n, \tag{2.5}$$

where $n=0,1,\cdots$. When $n=0$ the sums are empty and $R_0=F_\lambda(z,\alpha)$. The integration by parts procedure yields for R_n the representation

$$R_n=\int_0^\infty t^{\lambda-1}\exp\left(-z\left(\frac{t+\beta^2}{t}\right)\right)f_n(t)\,dt, \tag{2.6}$$

where f_n is defined by (2.4).

When a bound for $|f_n(t)|$ is available, say,

$$|f_n(t)|\leqq M_n, \qquad t\geqq 0, \quad n=0,1,\cdots, \tag{2.7}$$

then a bound for R_n reads

$$|R_n|\leqq M_n A_\lambda(z,\beta). \tag{2.8}$$

Since f_n depends on β, the quantity M_n may also depend on β. It follows that for bounded values of β, say $\beta \in [0, \beta_0]$, β_0 fixed and finite, the estimate (2.8) of the remainder R_n shows the asymptotic nature of the expansion (2.5), provided that (2.7) is satisfied.

We must point out that, in general, it is rather difficult to find realistic numbers M_n in order to obtain sharp estimates in (2.8). Also, the estimate in (2.7) is rather global, since it takes into account values of f_n in the complete interval $[0, \infty)$.

A sharper and more realistic bound for R_n may be obtained as follows. Let

$$w_\sigma(t) = \exp\{\sigma(t+\beta^2/t-2\beta)\}, \qquad t>0, \quad \sigma \geqq 0. \tag{2.9}$$

Observe that $w_\sigma(\beta)=1$ and that when $\sigma>0$

$$\lim_{t\to 0} w_\sigma(t) = \lim_{t\to+\infty} w_\sigma(t) = +\infty.$$

We assume that we can assign quantities σ_n and M_n, which may depend on β and which satisfy

$$\sigma_n \geqq 0, \qquad M_n \geqq 1+\varepsilon_n, \qquad \varepsilon_n \text{ fixed and positive,} \tag{2.10}$$

such that for all $t>0$ we have

$$|f_n(t)| \leqq M_n |f_n(\beta)| w_{\sigma_n}(t). \tag{2.11}$$

Then instead of (2.8) we obtain

$$|R_n| \leqq M_n |f_n(\beta)| \tilde{A}_\lambda(z,\beta), \qquad z>\sigma_n, \tag{2.12}$$

where

$$\tilde{A}_\lambda(z,\beta) = A_\lambda(z-\sigma_n,\beta)\, e^{-2\beta\sigma_n}. \tag{2.13}$$

When $f_n(\beta)=0$ a slight modification is needed. The idea about this approach is that in (2.11) function values outside a neighborhood of $t=\beta$ may be estimated very roughly, and that the integral, which results after inserting the right-hand side of (2.11) into (2.6), can be written in terms of one of the approximants in front of the series in (2.5).

A possible approach to computing M_n and σ_n of (2.11) is to start with trial values of M_n satisfying (2.10). Then we compute

$$\sigma_n = \sup_{t\geqq 0} \tilde{f}_n(t), \qquad \beta \text{ fixed in } [0,\infty),$$

where

$$\tilde{f}_n(t) = \frac{\ln|f_n(t)/[M_n f_n(\beta)]|}{t+\beta^2/t-2\beta}, \qquad t\neq\beta, \quad f_n(\beta)\neq 0.$$

Observe that the function defined in (2.13) satisfies

$$\tilde{A}_\lambda(z,\beta)/A_\lambda(z,\beta) = 1+o(1) \quad \text{as } z\to\infty,$$

uniformly with respect to $\beta \in [0,\infty)$. This follows from (2.2) and well-known asymptotic relations for the Bessel function.

3. Expansion at the internal saddle point. In the expansion (2.3) we have used function values of f at the negative saddle point $-\beta$. These values appear in the coefficients a_s, b_s of the expansion. The form of the expansion is very attractive, since only two special functions arise, and also since the parameters β and z are nicely

separated in both series. Although the expansion (2.3) has a canonical form, there remains the drawback that the function f must be defined at $(-\infty, 0]$ in order to obtain for β a uniformity domain $[0, \infty)$. For example, it is not possible to obtain such a uniformity domain when $f(t) = 1/(t+1)$. In this section we only expand the function f at the internal saddle point, and we formulate further conditions on f in order to obtain an optimal domain for β.

3.1. The functions $Q_s(\zeta)$ and $\tilde{Q}_s(\zeta)$. We expand f in the form

$$f(t) = \sum_{s=0}^{\infty} a_s(\beta)(t-\beta)^s, \qquad a_s = \frac{f^{(s)}(\beta)}{s!}. \tag{3.1}$$

Substituting (3.1) in (1.1), we obtain after interchanging the order of summation and integration the formal result

$$F_\lambda(z, \alpha) \sim z^{-\lambda} \sum_{s=0}^{\infty} a_s(\beta) Q_s(\zeta) z^{-s}, \qquad \text{as } z \to \infty, \tag{3.2}$$

where

$$Q_s(\zeta) = \zeta^{\lambda+s} \int_0^{\infty} t^{\lambda-1}(t-1)^s e^{-\zeta(t+1/t)}\, dt, \tag{3.3}$$

$$\zeta = \beta z. \tag{3.4}$$

The functions $Q_s(\zeta)$ can be expressed in terms of the modified Bessel functions defined in (1.2). It is easily verified that

$$Q_s(\zeta) = 2\zeta^{\lambda+s} \sum_{r=0}^{s} (-1)^{s-r} \binom{s}{r} K_{\lambda+r}(2\zeta). \tag{3.5}$$

On the other hand, integrating by parts in (3.3), we obtain the recursion relation

$$Q_{s+2} = (s+\lambda+1-2\zeta)Q_{s+1} + \zeta(2s+\lambda+1)Q_s + s\zeta^2 Q_{s-1}, \qquad s = 0, 1, 2, \cdots. \tag{3.6}$$

For proving the asymptotic properties of (3.2) it is useful to introduce the functions

$$\tilde{Q}_s(\zeta) = \zeta^{\lambda+s} \int_0^{\infty} t^{\lambda-1}|t-1|^s e^{-\zeta(t+1/t)}\, dt. \tag{3.7}$$

By applying Laplace's method it is found that for large positive values of ζ

$$\tilde{Q}_s(\zeta) \sim \zeta^{\lambda+(s-1)/2} e^{-2\zeta} \Gamma\left(\frac{s+1}{2}\right), \qquad s = 0, 1, 2, \cdots. \tag{3.8}$$

Furthermore, we have when z is fixed

$$\lim_{\beta \to 0} \tilde{Q}_s(\zeta) = \Gamma(\lambda+s).$$

3.2. Error bounds and interpretation of the expansion. We introduce a remainder in the expansion (3.2) by writing

$$f(t) = \sum_{s=0}^{n-1} a_s(\beta)(t-\beta)^s + R_n(t, \beta)(t-\beta)^n, \qquad n = 0, 1, 2, \cdots. \tag{3.9}$$

Then we obtain for (3.2)

$$F_\lambda(z, \alpha) = z^{-\lambda}\left[\sum_{s=0}^{n-1} a_s(\beta) Q_s(\zeta) z^{-s} + E_n(z, \alpha) z^{-n}\right], \tag{3.10}$$

where

$$E_n(z,\alpha)=z^{\lambda+n}\int_0^\infty t^{\lambda-1}(t-\beta)^n R_n(t,\beta)\exp\left(-z\left(\frac{t+\beta^2}{t}\right)\right)dt. \tag{3.11}$$

Let f be analytic in a connected domain Ω of the complex plane; Ω may depend on β, and we assume that the radius of convergence R_β of the expansion (3.1) satisfies the condition

$$R_\beta \geqq \rho(1+\beta)^\kappa, \qquad \beta\geqq 0 \qquad (\rho,\kappa \text{ fixed}, \rho>0, \kappa\geqq\tfrac{1}{2}). \tag{3.12}$$

This condition says that the distance between the singularities of f and the point $t=\beta$ should be of order $\mathcal{O}(\beta^\kappa)$, uniformly with respect to $\beta\in[0,\infty)$. When $\kappa<\frac{1}{2}$ the singularities of f are too close to the saddle point. Furthermore, we assume that f has the following growth condition in Ω: there is a real fixed number p such that

$$\sup_{t\in\Omega}(1+|t|)^{-p}|f(t)| \tag{3.13}$$

is bounded for $\beta\in[0,\infty)$.

The coefficients $a_s(\beta)$ of (3.1) can be written as

$$a_s(\beta)=\frac{1}{2\pi i}\int_{C_r}\frac{f(t)}{(t-\beta)^{s+1}}\,dt, \tag{3.14}$$

where C_r is a circle with centre β and radius $r(1+\beta)^\kappa$; r may depend on β, but should be uniformly bounded away from zero and small enough to keep C_r inside Ω. Using (3.14) we obtain the following form of Cauchy's inequality

$$|a_s(\beta)|\leqq r^{-s}M_r(\beta)(1+\beta)^{-s\kappa}, \tag{3.15}$$

where

$$M_r(\beta)=\sup_{t\in C_r}|f(t)|. \tag{3.16}$$

In the next theorem we introduce an asymptotic sequence $\{\phi_s\}$, which is constructed on the basis of the estimates in (3.7) and (3.15). For the concept of asymptotic scale and (generalized) asymptotic expansion we refer to [4, p. 25].

THEOREM 3.1. *Let* $\zeta=\beta z$, $\kappa\geqq\frac{1}{2}$, *and let*

$$\phi_s=\phi_s(z,\beta)=M_r(\beta)(1+\beta)^{-s\kappa}(1+\zeta)^{\lambda+(s-1)/2}\,e^{-2\zeta}z^{-s}, \qquad s=0,1,2,\cdots. \tag{3.17}$$

Then $\{\phi_s\}$ *is an asymptotic scale as* $z\to\infty$, *uniformly with respect to* $\beta\in[0,\infty)$.

Proof.

$$\frac{\phi_{s+1}}{\phi_s}=(1+\beta)^{-\kappa}\sqrt{\zeta+1}\,z^{-1}\leqq\frac{1}{\sqrt{z}} \quad \text{if } z\geqq 1. \tag{3.18}$$

Now we write the expansion (3.2) in the notation

$$z^\lambda F_\lambda(z,\alpha)\sim\sum_{s=0}^\infty a_s(\beta)Q_s(\zeta)z^{-s}; \qquad \{\phi_s\} \quad \text{as } z\to\infty, \tag{3.19}$$

and we have the following theorem.

THEOREM 3.2. *The expansion* (3.19) *is a uniform asymptotic expansion as* $z\to\infty$, *uniformly with respect to* $\beta\in[0,\infty)$.

Proof. According to the definition of generalized (uniform) asymptotic expansions, we have to prove

$$z^{-n}E_n(z,\beta)=\mathcal{O}(\phi_n),\qquad n=0,1,2,\cdots, \tag{3.20}$$

as $z\to\infty$, uniformly with respect to $\beta\in[0,\infty)$. The interval of integration in (3.11) is split up as follows

$$[0,\infty)=\Delta_-\cup[t_-,t_+]\cup\Delta_+, \tag{3.21}$$

where

$$\Delta_-=[0,t_-],\qquad \Delta_+=[t_+,\infty),\qquad t_\pm=\beta\pm r_1(1+\beta)^\kappa,\qquad 0<r_1<r,\quad r_1\text{ fixed}, \tag{3.22}$$

with r as in (3.14). When t_- happens to be negative, we replace it by 0. For $t\in[t_-,t_+]$ we can write

$$R_n(t,\beta)=\frac{1}{2\pi i}\int_{C_r}\frac{f(\tau)}{(\tau-t)(\tau-\beta)^n}\,d\tau, \tag{3.23}$$

with C_r as in (3.14). If $\tau\in C_r$, we have $|\tau-t|\geqq(r-r_1)(1+\beta)^\kappa$. Thus we obtain as in (3.15)

$$|R_n(t,\beta)|\leqq\frac{M_r(\beta)(1+\beta)^{-n\kappa}}{r^{n-1}(r-r_1)}. \tag{3.24}$$

Hence z^λ times the integral over $[t_-,t_+]$ in (3.11) gives a contribution which is bounded by

$$\begin{aligned}&\frac{z^\lambda M_r(\beta)(1+\beta)^{-n\kappa}}{r^{n-1}(r-r_1)}\int_{t_-}^{t_+}t^{\lambda-1}|t-\beta|^n\exp\left(-z\left(\frac{t+\beta^2}{t}\right)\right)dt\\&\quad=M_r(\beta)(1+\beta)^{-n\kappa}z^{-n}\tilde{Q}_n(\zeta)\mathcal{O}(1)\quad\text{as } z\to\infty,\end{aligned} \tag{3.25}$$

uniformly with respect to $\beta\in[0,\infty)$. Using (3.7), (3.8), and (3.17), we conclude that

$$z^{-n}E_n(z,\alpha)=I_-+I_++\mathcal{O}(\phi_n)\quad\text{as } z\to\infty, \tag{3.26}$$

uniformly with respect to $\beta\in[0,\infty)$, where $I_\pm$ are the contributions to (3.11) from $\Delta_\pm$. For $t\in\Delta_\pm$ we write

$$(t-\beta)^nR_n(t,\beta)=f(t)-\sum_{s=0}^{n-1}a_s(\beta)(t-\beta)^s,$$

and the proof is finished when we have shown that

$$z^\lambda\int_{\Delta_\pm}t^{\lambda-1}\,e^{-z(t+\beta^2/t)}\,g(t)\,dt=\mathcal{O}(\phi_n)\quad\text{as } z\to\infty, \tag{3.27}$$

uniformly with respect to $\beta\in[0,\infty)$, where $g(t)$ is $|f(t)|$ or $|a_s(t-\beta)^s|(0\leqq s\leqq n-1)$. In fact, it is possible to prove that

$$I_\pm\sim 0;\qquad\{\phi_s\}\quad\text{as } z\to\infty, \tag{3.28}$$

uniformly with respect to $\beta\in[0,\infty)$. That is, $I_\pm$ are asymptotically equal to zero with respect to the scale $\{\phi_s\}$. The proof of (3.28) is similar to that given for another type of integral in [5, Lemma 3.3] and will not be repeated here. □

The above theorem gives only an order estimate in terms of ϕ_n for the remainder defined in (3.11) and gives an interpretation of the asymptotic nature of the expansions (3.2) and (3.19). To obtain a numerical upper bound for $E_n(z,\lambda)$ we proceed as in

the previous section. Since $f(t)$ satisfies the growth condition (3.13), it is possible to find numbers M_n, σ_n satisfying (2.10), such that

$$|R_n(t,\beta)| \leqq M_n |a_n(\beta)| w_{\sigma_n}(t), \qquad 0<t<\infty. \tag{3.29}$$

Using this in (3.11), we obtain the bound

$$|E_n(z,\lambda)| \leqq M_n |a_n(\beta)|\, e^{-2\beta\sigma_n} \tilde{Q}_n(\zeta-\beta\sigma_n), \qquad z>\sigma_n. \tag{3.30}$$

When $a_n(\beta)$ happens to vanish as a function of $\beta \in [0,\infty)$, this approach needs a slight modification.

4. Application to confluent hypergeometric functions. We start with the confluent hypergeometric function defined by

$$\Gamma(a)U(a,b,x) = \int_0^\infty u^{a-1}(1+u)^{b-a-1}\, e^{-xu}\, du. \tag{4.1}$$

We consider a as the large parameter and x as a uniformity parameter in $[0,\infty)$; b is a fixed real parameter. We take $b \leqq 1$; the relation

$$U(a,b,x) = x^{1-b}U(a+1-b, 2-b, x)$$

can be used when $b>1$.

4.1. Transformation to the standard form. First we give a simple intermediate transformation. The function $[u/(u+1)]^a$ assumes its maximal value (on $[0,\infty)$) at $u=\infty$. This function controls the asymptotic behaviour of the integrand and, hence, we transform it to an exponential function by writing $u/(u+1)=\exp(-w)$. Then (4.1) becomes

$$\Gamma(a)U(a,1-\lambda,x) = \int_0^\infty w^{\lambda-1} \exp\left(-aw - \frac{x}{e^w-1}\right) \tilde{f}(w)\, dw, \tag{4.2}$$

where

$$\tilde{f}(w) = \left[\frac{1-e^{-w}}{w}\right]^{\lambda-1}. \tag{4.3}$$

We transform (4.2) into (1.1) with the help of the transformation

$$w + \frac{\nu}{e^w-1} = t + \frac{\beta^2}{t} + A, \tag{4.4}$$

where $\nu = x/a$ and β, A are to be determined. We compute them on the following condition on the mapping: the critical points of the w-function in (4.4) must correspond with the critical points of the t-function. Critical points are $\pm w_0$, $\pm t_0$, where

$$t_0 = \beta, \qquad w_0 = \cosh^{-1}(1+\nu/2) = \ln\left(1+\frac{\nu+W_0}{2}\right), \qquad W_0 = \sqrt{\nu^2+4\nu}. \tag{4.5}$$

It follows that

$$A = -\frac{\nu}{2}, \qquad \beta = \frac{w_0+\sinh w_0}{2} = \frac{1}{2}\ln\left(1+\frac{\nu+W_0}{2}\right) + \frac{1}{4}W_0. \tag{4.6}$$

From the simple differential equation

$$\frac{d\beta}{d\nu} = \frac{1}{4}\sqrt{(\nu+4)/\nu}$$

and a Taylor expansion of the right-hand side, it follows that β^2 of (4.4) is an analytic function of ν, at least in the disc $|\nu|<4$. Conversely, ν is an analytic function of β^2 in some neighborhood of the origin. These domains can be extended to domains containing $[0,\infty)$.

With these values of A, β the mapping $w\mapsto t$ is regular at $w=\pm w_0$ and at $w=0$. In fact it is regular in $\mathbb{R}$ and as a conformal mapping in a large domain Ω of the complex plane. We have the correspondences

$$t(\pm\infty)=\pm\infty, \qquad t(\pm w_0)=\pm\beta, \qquad t(0)=0. \tag{4.7}$$

More details on the mapping are given in the next subsection.

Using transformation (4.4) in (4.2), we arrive at the standard form

$$F_\lambda(z,\alpha)=\Gamma(a)\,e^{-x/2}U(a,b,x)=\int_0^\infty t^{\lambda-1}\,e^{-zt-\alpha/t}f(t)\,dt, \tag{4.8}$$

with $z=a$, $\alpha=z\beta^2$, $\lambda=1-b$, β defined in (4.6) with $\nu=x/a$, and

$$f(t)=\left(\frac{1-e^{-w}}{t}\right)^{\lambda-1}\frac{dw}{dt}, \qquad \frac{dw}{dt}=\left(\frac{e^w-1}{t}\right)^2\frac{t^2-\beta^2}{(e^w-1)^2-\nu e^w}. \tag{4.9}$$

The function $t(w)$ defined in (4.4) is an odd function of w. This easily follows from rewriting (4.4) in the form

$$\frac{1}{2}\nu+w+\frac{\nu}{e^w-1}=t+\frac{\beta^2}{t}. \tag{4.10}$$

After these preparations the expansion of (3.2) can be constructed. The expansion holds uniformly with respect to $\beta\in[0,\infty)$; that is, uniformly with respect to $x\in[0,\infty)$.

The asymptotic nature of the expansion follows from combining (3.20) and (3.17). For this particular case we can derive an upper bound for $M_r(\beta)$ of (3.16). The t-values on the circle are written as $t=\beta+\tau\sqrt{\beta+1}$, with $|\tau|=r$, r fixed. When β and ν are large, we derive from (4.6) $\beta=\nu/4+\ln\sqrt{\nu}+\mathcal{O}(1)$. So, for large values of β, we obtain (using (4.10)) $t+\beta^2/t-\nu/2=w+\nu/(e^w-1)=\ln\nu+\tau^2+\mathcal{O}(1)$. That is, $w\sim\ln\nu$. Then it follows from (4.9) that $f(t)=\mathcal{O}(\beta^{1/2-\lambda})$, $t\in C_r$. Consequently, we can find a fixed number K, such that

$$M_r(\beta)\leqq K(\beta+1)^{1/2-\lambda}, \qquad \beta\in[0,\infty). \tag{4.11}$$

To conclude this subsection, we give the first coefficient $a_0(\beta)$ of (3.2). A few calculations based on (4.9) and l'Hôpital's rule yield

$$\left.\frac{dw}{dt}\right|_{t=\pm\beta}=\sqrt{2\tanh(w_0/2)/\beta}.$$

So we obtain

$$a_0(\beta)=\sqrt{2\tanh(w_0/2)/\beta}\left(\frac{1-e^{-w_0}}{\beta}\right)^{\lambda-1}. \tag{4.12}$$

4.2. Analytical properties of the mapping (4.4). We now consider the mapping (4.4) in more detail. We restrict w to the strip

$$H=\{w\,|\,\Re w\in\mathbb{R},\,\Im w\in[-\pi,\pi]\}, \tag{4.13}$$

and we prove the following.

THEOREM 4.1. *Let Ω be the image of H under the mapping $w\mapsto t$ defined in* (4.4). *Let $\nu\in[0,\infty)$ and let A, β be defined by* (4.6). *Then $t(w,\beta)$ is analytic in H.*

In the following proof we show that $t(w,\beta)$ and $w(t,\beta)$ are analytic in a fixed neighborhood of $(0,0)$. Accordingly, we concentrate on small (complex) values of the parameters. For remaining values the proof is much easier. For instance, when β is bounded away from zero, the critical points $t=\pm\beta$ and the pole at $t=0$ of the right-hand side of (4.4) are well separated. The preparations for applying the Implicit Function Theorem mentioned below are more straightforward then.

Proof. From (4.6) it follows that

$$w_0=\beta[1+o(1)], \qquad \nu=\beta^2[1+o(1)] \quad \text{as } \beta\to 0.$$

Recall that $t(w)$ is an odd function of w (see (4.10)). We introduce a function $y=y(w,\beta)$ by writing

$$t=w\left[\frac{\beta}{w_0}+(w^2-w_0^2)y\right]. \tag{4.14}$$

This matches the points $w=0\Leftrightarrow t=0$ and also the critical points $w=\pm w_0\Leftrightarrow t=\pm\beta$; y is an even function of w and should vanish with β. Substituting (4.14) in (4.10), we obtain

$$\frac{\nu t}{w}\frac{\phi(w)-\phi(w_0)}{w^2-w_0^2}+\frac{\beta}{w_0}\left(1-\frac{\beta}{w_0}\right)+(w^2-w_0^2)\left(1-\frac{2\beta}{w_0}\right)y-w^2(w^2-w_0^2)^2y^2=0,$$

where $\phi(w)=w/(\exp w-1)-1+w/2$. We expand

$$\frac{\phi(w)-\phi(w_0)}{w^2-w_0^2}=\sum_{s=1}^{\infty}b_s(w^2-w_0^2)^{s-1}, \qquad b_1=\frac{\beta-w_0}{\nu w_0}.$$

Since $\phi(w)$ is analytic if $|w|<2\pi$, the series converges if β and w are small. Finally, we obtain the equation $F(y,w,\beta)=0$, where F is given by

$$\frac{\nu\beta}{w_0}\sum_{s=2}^{\infty}b_s(w^2-w_0^2)^{s-2}+\nu y\sum_{s=1}^{\infty}b_s(w^2-w_0^2)^{s-1}+\left(1-\frac{2\beta}{w_0}\right)y-w^2y^2.$$

The series represents analytic functions of w, w_0. When β is small, w_0 is an analytic function of β (see (4.6)). Hence, F is analytic in a fixed neighborhood of $(0,0,0)$, $F(0,0,0)=0$, and $F_y(0,0,0)=-1$. After these preparations we can use an Implicit Function Theorem (see, for instance, [1, p. 36]) and solve for $y(w,\beta)$; it is analytic in a fixed neighborhood of $(0,0)$. By using (4.14) it follows that the same holds for $t(w,\beta)$. □

The first terms in the expansion

$$t(w,\beta)=c_1(\beta)w+c_3(\beta)w^3+\cdots$$

easily follow from (4.10). We have

$$c_1(\beta)=\frac{\beta^2}{\nu}, \qquad c_3(\beta)=\frac{c_1^2+\dfrac{\beta^2}{6}-c_1-\dfrac{\nu c_1}{4}}{\nu}.$$

THEOREM 4.2. *The mapping* (4.4) *is univalent in H.*

Proof. First we show that the mapping is univalent on

$$\mathscr{L}_+=\{w=u+iv\,|\,u\in\mathbb{R},\ v=\pi\},$$

which is the upper part of the boundary ∂H of H. We write $t=re^{i\theta}$. The image of $\mathscr{L}_+$ in the t-plane is defined by the equations

$$\pi=r\sin\theta\left(1-\frac{\beta^2}{r^2}\right), \quad \Psi(u)=\Phi(\theta), \tag{4.15}$$

where

$$\Psi(u)=\nu/2+u-\nu/(e^u+1), \qquad \Phi(\theta)=r\cos\theta\left(1+\frac{\beta^2}{r^2}\right).$$

The first equation in (4.15) defines a curve given by

$$r(\theta)=\frac{\pi+\sqrt{\pi^2+4\beta^2\sin^2\theta}}{2\sin\theta}, \qquad 0<\theta<\pi.$$

It follows that $r>\pi/\sin\theta$. Furthermore, we have

$$\Psi(-\infty)=\Phi(\pi)=-\infty, \qquad \Psi(+\infty)=\Phi(0)=+\infty.$$

The function $\Psi(u)$ is one-to-one on $\mathbb{R}$. The same is true for $\Phi(\theta)$ on $(0,\pi)$, but the proof requires a little extra work. We have, using the first equation in (4.15),

$$\frac{dr}{d\theta}=-\frac{r\cos\theta(r^2-\beta^2)}{\sin\theta(r^2+\beta^2)}.$$

It follows that

$$\frac{d\Phi(\theta)}{d\theta}=-\frac{\sin\theta(r^2+\beta^2)}{r^3}\left[r^2+\left(\frac{dr}{d\theta}\right)^2\right], \tag{4.16}$$

which shows that $\Phi(\theta)$ is one-to-one on $(0,\pi)$. We infer that for each value $u\in\mathbb{R}$ we can find one and only one value $\theta\in(0,\pi)$, such that $\Psi(u)=\Phi(\theta)$, and, hence, one and only one value $r(\theta)$. Since $t(w)$ is an odd function of w (see (4.10)), the mapping $w\mapsto t$ is one-to-one on ∂H. When $\Re w\to\pm\infty$ we have $w\sim t$. Hence the mapping $t(w)$ is also one-to-one as $w\to\infty$, $w\in H$. We now consider a large closed rectangle $ABCD$ of which upper side AB and lower side CD are finite parts of ∂H, and BC and AD are far away to the right and to the left, respectively. From the above arguments it is not difficult to conclude that the mapping is univalent on BC, AD, and on the whole Jordan curve $ABCDA$, provided that the vertical sides are far away. Then we use a well-known result of complex function theory, which says that consequently the mapping is also univalent in the interior of rectangle $ABCD$, since it is analytic there. See [3, Vol. II, p. 118]. We can take the finite rectangle as large as we please. Thus the result also holds for H. □

For the uniform expansion of (4.8) we take Ω as the image of the strip H under the mapping $w\mapsto t$. From $f(t)$ defined in (4.9) it follows that (3.13) is bounded in Ω if $p=1-\lambda$ and that $M_r(\beta)$ of (3.16) is well defined. There remains to show that the radius of convergence R_β of the series in (3.1) satisfies (3.12). It appears that we must take $\kappa=\frac{1}{2}$. In fact, we show that Ω contains a disc around β with radius $\rho\sqrt{\beta+1}$ (ρ fixed), for all $\beta\geqq 0$. The points of intersection of the circle with radius r around the point $t=\beta$ with the curve defined by the first equation of (4.15) are governed by the equations (we write $t=\sigma+i\tau$)

$$(\sigma-\beta)^2+\tau^2=r^2, \qquad \pi=\tau\left(1-\frac{\beta^2}{\sigma^2+\tau^2}\right), \qquad \tau>\pi.$$

When we require that the circle is tangent at the curve we have the extra condition

$$\frac{\sigma-\beta}{\tau}=\frac{2\sigma(\tau-\pi)^2}{\pi\beta^2+2\tau(\tau-\pi)^2}.$$

This equation is obtained by equating $d\tau/d\sigma$ of both equations and eliminating $\tau^2+\sigma^2$ by using the second one. For large values of β the solution of these three equations reads

$$\tau=\pi+a\sqrt{\beta}[1+o(1)], \qquad \sigma=\beta+b\sqrt{\beta}[1+o(1)], \qquad r=c\sqrt{\beta}[1+o(1)],$$

with $a=b=\sqrt{\pi/2}$, $c=\sqrt{\pi}$.

This shows that Ω is large enough to apply Theorem 3.2. From a further analysis it follows that the value $\kappa=\frac{1}{2}$ is best possible in this case. Apart from the real critical points $\pm w_0$ given in (4.5), which are regular points for the mapping, we have other ones located at $\pm w_0\pm 2\pi ni$, $n=1, 2, \cdots$. For large values of β those are mapped at a distance $\mathcal{O}(\sqrt{\beta})$ from the critical point $t=\beta$.

Remark 4.1. The behaviour of $f(t)$ of (4.9) in the left half-plane $\Re t<0$ is quite different from that in $\Re t>0$, except when $\lambda=1$. Consequently, the approach of § 2 is less attractive. See also Remark 2.1.

Remark 4.2. When $b=\frac{1}{2}$, (4.1) is a parabolic cylinder function, and the functions $Q_s(\zeta)$ defined in (3.3), (3.5) are elementary functions ($\lambda=\frac{1}{2}$). Then (3.2) gives an expansion of the parabolic cylinder function $D_\nu(z)$, as $\nu\to-\infty$, which is uniformly valid with respect to $z\in[0,\infty)$.

5. A second uniformity parameter. In this section we consider (1.1) with λ as a second uniformity parameter in $[0,\infty)$. Thus we take further advantage of the fact that the modified Bessel function is a function of two variables. In this case it is convenient to put the reciprocal gamma function in front of the integral. So, now we write

$$F_\lambda(z,\alpha)=\frac{1}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}\,e^{-zt-\alpha/t}f(t)\,dt. \tag{5.1}$$

In [5] we considered (5.1) with $\alpha=0$, again with $z\to\infty$ and λ as a uniformity parameter in $[0,\infty)$. In [6] we applied the present method for a loop integral (without proofs) to the case of Laguerre polynomials.

We write $\lambda=\mu z$. The critical points of the integrand are now defined as the points where the derivative of $t+\beta^2/t-\mu\ln t$ vanishes. This gives the real critical points

$$t_\pm=\frac{\mu\pm T}{2}, \qquad T=\sqrt{\mu^2+4\beta^2}. \tag{5.2}$$

Observe that also in this case one of the real saddle points is outside the interval of integration, and that the "phase function" that is used to compute the critical points has a logarithmic singularity at $t=0$. The two critical points coincide with this singularity when β and μ both vanish. At the same moment, however, the logarithmic singularity disappears.

First we construct an expansion by using the integration by parts procedure of § 2. The modification of (2.1) is

$$f(t)=c_0+d_0(t-t_+)+(t-\mu-\beta^2/t)h(t). \tag{5.3}$$

Using this in (5.1) we obtain, after repeating the procedure,

$$F_\lambda(z,\alpha)=C(z,\beta,\mu)\sum_{s=0}^{n-1}c_sz^{-s}+D(z,\beta,\mu)\sum_{s=0}^{n-1}d_sz^{-s}+z^{-n}R_n. \tag{5.4}$$

The functions in front of the series are again combinations of Bessel functions as in (2.2). We have

$$C(z,\beta,\mu)=\frac{2\beta^\lambda}{\Gamma(\lambda)}K_\lambda(2\beta z), \qquad D(z,\beta,\mu)=\frac{2\beta^\lambda}{\Gamma(\lambda)}[\mu K_{\lambda+1}(2\beta z)-t_+K_\lambda(2\beta z)].$$

The coefficients c_s, d_s follow from the recursion relation

$$f_0(t)=f(t), \qquad f_s(t)=t\frac{d}{dt}h_{s-1}(t)=c_s+d_s(t-t_+)+\left(t-\mu-\frac{\beta^2}{t}\right)h_s(t),$$

$$c_s=f_s(t_+), \qquad d_s=\frac{f(t_+)-f(t_-)}{t_+-t_-}.$$

The remainder R_n in (5.4) can be written in the form

$$R_n=\frac{1}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}\,e^{-zt-\alpha/t}f_n(t)\,dt.$$

A bound can be constructed by using constants σ_n, M_n satisfying (2.10), and using a function

$$w_\sigma(t)=\exp\left(\sigma\left(t+\frac{\beta^2}{t}-\mu\ln t-t_+-\frac{\beta^2}{t_+}+\mu\ln t_+\right)\right)$$

such that, as in (2.11), for all $t>0$

$$|f_n(t)|\leqq M_n|f_n(t_+)|w_{\sigma_n}(t).$$

Then we obtain

$$|R_n|\leqq M_n|f_n(t_+)|\tilde{C}(z,\beta,\mu), \qquad z>\sigma_n,$$

where

$$\tilde{C}(z,\beta,\mu)=C(z-\sigma_n,\beta,\mu)\exp\left(-\sigma_n\left(t_++\frac{\beta^2}{t_+}-\mu\ln t_+\right)\right).$$

When $f_n(t_+)=0$, a slight modification is needed. An optimal value of σ_n follows from the method described in § 2.2.

The analogue of the expansion of § 3 is obtained by substituting

$$f(t)=\sum_{s=0}^{n-1}c_s(\beta,\mu)(t-t_+)^s+R_n(t,\beta,\mu)(t-t_+)^n, \qquad c_s=\frac{f^{(s)}(t_+)}{s!}. \tag{5.5}$$

So we obtain

$$F_\lambda(z,\alpha)=z^{-\lambda}\left[\sum_{s=0}^{n-1}c_s(\beta,\mu)P_sz^{-s}+E_n(z,\alpha,\lambda)z^{-n}\right], \tag{5.6}$$

where

$$E_n(z,\alpha,\lambda)=z^{\lambda+n}\int_0^\infty t^{\lambda-1}(t-t_+)^nR_n(t,\beta,\mu)\exp\left(-z\left(\frac{t+\beta^2}{t}\right)\right)dt,$$

$$P_s=\frac{z^{\lambda+s}}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}(t-t_+)^s\exp\left(-z\left(\frac{t+\beta^2}{t}\right)\right)dt$$

$$=\frac{2z^{\lambda+s}\beta^\lambda}{\Gamma(\lambda)}\sum_{r=0}^{s}\binom{s}{r}(-t_+)^{s-r}\beta^rK_{\lambda+r}(2\beta z).$$

A recursion relation for P_s follows from the above integral representation.

$$\tilde{P}_s=\frac{z^{\lambda+s}}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}|t-t_+|^s\exp\left(-z\left(\frac{t+\beta^2}{t}\right)\right)dt$$

$$\sim\frac{\eta^{\lambda+(s-1)/2}}{\Gamma(\lambda)}\exp\left(-\eta\left(1+\frac{\beta^2}{t_+^2}\right)\right)\left[\frac{\beta^2+t_+^2}{2t_+^2}\right]^{-(s+1)/2}\Gamma\left(\frac{s+1}{2}\right) \quad \text{as } \eta\to\infty, \tag{5.7}$$

where $\eta = zt_+$. Since z is the large parameter, η is large if at least one of the uniformity parameters β, μ is bounded away from zero.

The coefficients c_s and the remainder R_n can be written as

$$c_s(\beta,\mu)=\frac{1}{2\pi i}\int_{C_r}\frac{f(\tau)}{(\tau-t_+)^s}\,d\tau, \qquad R_n(t,\beta,\mu)=\frac{1}{2\pi i}\int_{C_r}\frac{f(\tau)}{(\tau-t)(\tau-t_+)^n}\,d\tau,$$

where C_r is a circle around t_+ with radius $r(1+t_+)^\kappa$, $\kappa \geqq \frac{1}{2}$, $r>0$. We accept that f depends on both uniformity parameters β, μ, and we assume that the domain of analyticity Ω is large enough to contain such a circle for all β, $\mu \geqq 0$.

As in § 3 we have the following theorems. The quantity $M_r(\beta,\mu)$ is defined as in (3.16); we also assume that (3.13) is bounded for all β, $\mu \in [0,\infty)$.

THEOREM 5.1. *Let $\eta = zt_+$, $\kappa \geqq \frac{1}{2}$, and let for $s = 1, 2, \cdots$*

$$\chi_s=\frac{M_r(\beta,\mu)}{\Gamma(\lambda)z^s}\,\frac{(1+\eta)^{(s-1)/2}}{(1+t_+)^{s\kappa}}\exp\left(-\eta\left(1+\frac{\beta^2}{t_+^2}\right)\right)\left[\frac{\beta^2+t_+^2}{2t_+^2}\right]^{-(s+1)/2}. \tag{5.8}$$

Then $\{\chi_s\}$ is an asymptotic scale as $z\to\infty$, uniformly with respect to β, $\mu \in [0,\infty)$.

THEOREM 5.2. *The expansion*

$$z^\lambda F_\lambda(z,\alpha)\sim\sum_{s=0}^{\infty}c_s(\beta,\mu)P_s z^{-s}; \qquad \{\chi_s\} \quad \text{as } z\to\infty, \tag{5.9}$$

is a uniform asymptotic expansion as $z\to\infty$, uniformly with respect to β, $\mu \in [0,\infty)$.

A bound for the remainder E_n of (5.6) can be constructed by combining the methods used for (3.30) and the above estimate for the remainder of (5.4).

5.1. Application to a confluent hypergeometric function. Our starting point is (cf. (4.2))

$$\frac{\Gamma(a)}{\Gamma(\lambda)}U(a,1-\lambda,x) =\frac{1}{\Gamma(\lambda)}\int_0^\infty \exp\left(-z\left[-\mu\ln(1-e^{-w})+w+\frac{\nu}{e^w-1}\right]\right)\frac{dw}{1-e^{-w}}, \tag{5.10}$$

with $z=a$, $\mu=\lambda/z$, $\nu=x/z$. The real critical points of the "phase function" are

$$w_\pm=\ln\left(1+\frac{\mu+\nu\pm W}{2}\right), \qquad W=\sqrt{(\mu+\nu)^2+4\nu}. \tag{5.11}$$

The transformation to the standard form (5.1) reads

$$-\mu\ln(e^w-1)+(\mu+1)w+\frac{\nu}{e^w-1}=t+\frac{\beta^2}{t}-\mu\ln t+A; \tag{5.12}$$

A, β are determined by substituting $w_\pm$ and $t_\pm$, where $t_\pm$ are the critical points defined in (5.2). We have the correspondences

$$t(\pm\infty)=\pm\infty, \qquad t(w_\pm)=t_\pm, \qquad t(0)=0.$$

Observe that the introduction of a second parameter (here in the form of μ) does not require a third constant in the equation (5.12). It has the same number of constants as (4.4). In fact, in order to obtain a regular mapping $w\mapsto t$, the constants multiplying the log-functions in the left- and right-hand side of (5.12) must be the same. We assume that the log-functions take their principal branches.

Elimination of A from the two equations (5.12) (with $w=w_{\pm}$, $t=t_{\pm}$) gives a relation for the unknown parameter β in terms of μ, ν:

$$(\mu+1)\ln\frac{2+\mu+\nu+W}{2+\mu+\nu-W}-\mu\ln\frac{W+\mu+\nu}{W-\mu-\nu}+W=2T-\mu\ln\frac{T+\mu}{T-\mu}. \tag{5.13}$$

By considering $\mu\in[0,\infty)$ as a fixed parameter, we obtain a more transparent relation for $\beta(\nu)$ in the form of a differential equation:

$$\frac{d\beta(\nu)}{d\nu}=\frac{\beta W}{2\nu T},\qquad \beta(0)=0. \tag{5.14}$$

The value of A follows from (5.12) by substituting $w=w_+$, $t=t_+$. We have

$$A=(\mu+1)w_+-\mu\ln\frac{\mu+\nu+W}{\mu+T}-\frac{\mu+\nu-W}{2}-T.$$

Using (5.13), we can eliminate $W/2-T$ and we obtain

$$A=\frac{1}{2}\left[(\mu+1)\ln(\mu+1)+\mu\ln\frac{\beta^2}{\nu}-\mu-\nu\right]. \tag{5.15}$$

The transformation (5.12) is discussed in the next subsection. By using it in (5.10) we obtain the standard form (5.1):

$$F_\lambda(z,\alpha)=\frac{e^{zA}\Gamma(a)}{\Gamma(\lambda)}U(a,1-\lambda,x)=\frac{1}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}e^{-zt-\alpha/t}f(t)\,dt,$$

where $z=a$, $\alpha=z\beta^2$; β^2 follows from (5.13) with $\mu=\lambda/z$, $\nu=x/z$. Furthermore,

$$f(t)=\frac{t}{1-e^{-w}}\frac{dw}{dt}=\frac{e^w(e^w-1)}{t}\,\frac{t^2-\mu t-\beta^2}{(e^w-1)^2-(\mu+\nu)(e^w-1)-\nu}. \tag{5.16}$$

The first coefficient of (5.9) equals $f(t_+)$. A few computations give

$$c_0(\beta,\mu)=e^{w_+/2}\sqrt{T/W}.$$

The function f satisfies $f(t)\sim t$ as $t\to+\infty$, whereas f is exponentially small at $-\infty$. This time we can also derive an expansion based on (5.4).

5.2. Analytical properties of the mapping (5.12). The mapping $w\mapsto t$ defined in (5.12) is one-to-one on the strip H given in (4.13). First we prove this property for the boundary. The proof is similar to that for Theorem 4.2. The equations for the image of the upper part of ∂H are given by (cf. (4.15))

$$\pi=r\sin\theta\left(1-\frac{\beta^2}{r^2}\right)-\mu\theta,\qquad \Psi(u)=\Phi(\theta),$$

where

$$\Psi(u)=-A+(\mu+1)u-\frac{\nu}{e^u+1},\qquad \Phi(\theta)=r\cos\theta\left(1+\frac{\beta^2}{r^2}\right)-\mu\ln r.$$

It follows that the image is given by

$$r(\theta)=\frac{\mu\theta+\pi+\sqrt{(\mu\theta+\pi)^2+4\beta^2\sin^2\theta}}{2\sin\theta},\qquad 0<\theta<\pi.$$

The function $\Psi(u)$ is one-to-one on $\mathbb{R}$. When we compute $d\Phi(\theta)/d\theta$, we find the same expression as in (4.16). As in Theorem 4.2, we conclude that the mapping is univalent on the boundary for all β, $\mu\in[0,\infty)$.

It remains to show that the mapping is analytic inside H. The interesting question is: Is $t(w)$ analytic at $t=0$, $t=w_{\pm}$, uniformly with respect to the parameters ν, μ? Especially interesting are small values of the parameters, since then the critical points coalesce with the pole and log-singularity at $w=0$. When one of the parameters is bounded away from the origin, the critical points $w_{\pm}$ are well separated. In that case the problem is simpler. Here we prove that $t(w, \nu, \mu)$ is analytic for complex values of the three arguments in a fixed neighborhood of $(0, 0, 0)$. The proof follows the idea of § 4.2.

First we have the following theorem.

THEOREM 5.1. *$\beta^2=\beta^2(\nu, \mu)$ defined by (5.13) is an analytic function of ν, μ.*

Proof. As remarked earlier, we concentrate on small values of the parameters. For $\mu=0$ the relation between ν and β is given in (4.6), and we have mentioned there that $\beta^2(\nu, 0)$ is analytic in the domain of interest. On the other hand, we have the expansion

$$\beta^2(\nu, \mu) \sim c_1(\mu)\nu + c_2(\mu)\nu^2 + c_3(\mu)\nu^3 + \cdots \quad \text{as } \nu \to 0.$$

The coefficients c_s are analytic functions of μ. The first few easily follow from (5.13):

$$c_1(\mu) = e^{(\mu+1)\ln(\mu+1)/\mu - 1} = 1 + \frac{1}{2}\mu - \frac{1}{24}\mu^2 + \mathcal{O}(\mu^3) \quad \text{as } \mu \to 0,$$

$$c_2(\mu) = \frac{c_1(\mu)[\mu + 2 - 2c_1(\mu)]}{\mu^2} = \frac{1}{12} + \mathcal{O}(\mu) \quad \text{as } \mu \to 0.$$

Next we observe that the quantity T of (5.2) is singular at $\beta^2=-\mu^2/4$ and that W of (5.11) has singular points at $\nu=\nu_0$, $\nu=\nu_1$, where

$$\nu_0 = -(\mu+2) + 2\sqrt{\mu+1}, \qquad \nu_1 = -(\mu+2) - 2\sqrt{\mu+1}. \tag{5.17}$$

It is obvious that the singularities at $-\mu^2/4$, ν_0 must correspond. That is, a necessary condition for β^2 to be regular for small values of $|\mu|$ is $\beta^2(\nu_0, \mu) = -\mu^2/4$. Note that $\nu_0 \sim -\mu^2/4$ as $\mu \to 0$ and that (5.13) is satisfied when we substitute $T=W=0$.

We "remove" the singularity at $\nu=\nu_0$ from (5.13), and we introduce a function $X=X(q, \mu)$ by writing

$$\frac{T+\mu}{T-\mu}\,\frac{W-\mu-\nu}{W+\mu+\nu} = \frac{1+\sqrt{q}\,X}{1-\sqrt{q}\,X}, \qquad q = \nu - \nu_0. \tag{5.18}$$

In other words,

$$\sqrt{q}\,X = \frac{\mu W - (\nu+\mu)T}{WT - \mu(\mu+\nu)}, \qquad T = \mu[W + (\nu+\mu)\sqrt{q}\,X]/D, \tag{5.19}$$

$$D = \nu + \mu + \sqrt{q}\,XW.$$

Now we can rewrite (5.13) in the form $K+L+M=0$, with

$$K = (W-2T)D = W(\nu-\mu) + \sqrt{q}\,X(\nu^2 + 4\nu - \mu^2),$$

$$L = D(\mu+1)\ln\frac{2+\mu+\nu+W}{2+\mu+\nu-W} = D(\mu+1)\ln\frac{1+\sqrt{q}\,Z}{1-\sqrt{q}\,Z},$$

$$M = D\mu\ln\frac{T+\mu}{T-\mu}\,\frac{W-\mu-\nu}{W+\mu+\nu} = D\mu\ln\frac{1+\sqrt{q}\,X}{1-\sqrt{q}\,X},$$

where

$$Z=\frac{\sqrt{\nu-\nu_1}}{2+\mu+\nu}=\frac{\sqrt{q+\nu_0-\nu_1}}{q+2+\mu+\nu_0}.$$

We expand $K+L+M$ in powers of q. A first observation is that $F(q, X, \mu):=(K+L+M)/\sqrt{q}$ is a function of q, X, μ, the factor $\sqrt{q}$ being completely removed. We expand F in powers of q. We have

$$F(q, X, \mu)=F_0+F_1q+F_2q^2+\cdots,$$

where $F_s(X, \mu)$ do not explicitly depend on q (or ν). We compute

$$F_0=(\nu_0-\mu)\sqrt{\nu_0-\nu_1}-2\mu(\nu_0+\mu)X+\frac{2\sqrt{\nu_0-\nu_1}(\nu_0+\mu)(\mu+1)}{2+\mu+\nu_0}+2\mu(\nu_0+\mu)X.$$

It appears that $F_0(X, \mu)\equiv 0$, and that, hence, we can continue with the equation $G(q, X, \mu):=F/q=F_1+F_2q+\cdots=0$. We claim that the equation $G(q, X, \mu)=0$ can be solved for $X=X(q, \mu)$, and that X is analytic for small values of both arguments. By calculating some limits, it follows from (5.18) or (5.19) that $X(0, 0)=-\frac{1}{2}$. This is used to show that $G(0, -\frac{1}{2}, 0)=F_1(0, 0)=0$. In order to apply an Implicit Function Theorem (see [1, p. 36]), we need to show that G is analytic in a neighborhood of $(0, -\frac{1}{2}, 0)$ and that $G(0, -\frac{1}{2}, 0)=0$, $G_X(0, -\frac{1}{2}, 0)\neq 0$. It is straightforward to verify that $G(q, X, \mu)$ is analytic in a neighborhood of $(0, -\frac{1}{2}, 0)$. Furthermore, $G_X(0, -\frac{1}{2}, 0)=\partial F_1/\partial X=4$ at $(X, \mu)=(-\frac{1}{2}, 0)$. We have shown that we can solve the equation $G=0$ and that the solution $X(q, \mu)$ is analytic in a fixed neighborhood of $(0, 0)$.

It remains to show that β^2 is analytic. We consider T of (5.2) given in the middle of (5.19). We are done when we have shown that μ/D is bounded away from zero when μ is small, since then we can divide the denominator of T by μ. From the above result it follows that we can expand

$$X(q, \mu)=X_0(q)+X_1(q)\mu+\cdots,$$

where the coefficients X_s are analytic functions of q. From the first equation of (5.19) we compute $X_0=-1/\sqrt{\nu-\nu_1}=-1/\sqrt{\nu+4}$. Hence

$$D=\nu+\mu+(\nu-\nu_0)\sqrt{\nu-\nu_1}\,X_0+\mathcal{O}(\mu)=\mathcal{O}(\mu)$$

as $\mu\to 0$. It now follows that T^2 is an analytic function of q, μ in a fixed neighborhood of $(0, 0)$, and, consequently, that β^2 is analytic. This proves the theorem. □

Remark 5.1. It is possible to base a proof on the differential equation (5.14). The condition $\beta(0)=0$ is not enough to prove the theorem, since the ratio β^2/ν (at $\nu=0$) turns out to be undefined. Requiring that this ratio equals $c_1(\mu)$ is sufficient, however.

In Theorem 4.1 we expanded the functions of (4.10) at the critical points $\pm w_0$, and in (4.14) we used a representation of t in which y can be viewed as a part of the complete expansion. In fact, (4.14) is a change of variables. In the present case we expand at the critical points $w_\pm$, and the expansions have the form

$$\psi(w)=\sum_{k=0}^{\infty}[a_k+wb_k]V^k, \qquad V=V(w)=(w-w_-)(w-w_+). \tag{5.20}$$

When ψ is sufficiently smooth, the coefficients a_k, b_k are uniquely defined. The first

few are given by

$$a_0=\frac{\psi_-w_+-\psi_+w_-}{w_+-w_-}, \qquad b_0=\frac{\psi_+-\psi_-}{w_+-w_-},$$

$$a_1=\frac{b_0w_0-\psi'_+w_--\psi'_-w_+}{(w_+-w_-)^2}, \qquad b_1=\frac{\psi'_++\psi'_--2b_0}{(w_+-w_-)^2},$$

where $w_0=w_++w_-$, and $\psi_+=\psi(w_+)$, etc. For analytic functions the coefficients can be represented as Cauchy-type integrals. We have

$$(5.21)\quad a_k=\frac{1}{2\pi i}\int_C (w-w_0)V^{-k-1}(w)\psi(w)\,dw, \qquad b_k=\frac{1}{2\pi i}\int_C V^{-k-1}(w)\psi(w)\,dw,$$

where C is a contour around the two critical points; ψ must be analytic inside C and continuous on C. This can be verified by substituting a new variable $w=v+w_0/2$. Then we have

$$\psi(w)=f\left(v+\frac{w_0}{2}\right)=\sum c_kV^k+v\sum b_kV^k, \qquad c_k=a_k+\frac{1}{2}w_0b_k.$$

By separating odd and even parts (with respect to v), and representing c_k, b_k as Cauchy integrals in the V-plane, we arrive at (5.21). (Note that a circle around the origin in the w-plane is traversed twice in the V-plane.) For MacLaurin series the domain of convergence is a disc. For expansions as in (5.20) the domain of convergence is defined by $|V(w)|<|V(w_s)|$, where w_s is a singularity of ψ; this domain is bounded by a Cassini's oval with foci at $w_\pm$. See also [7, Exercise 24, p. 149].

The parameter t of (5.12) is represented in the form

$$(5.22)\qquad t=w[B+Cw+V(w)y],$$

where B, C do not depend on w, and we require that the points $\{w_-, 0, w_+\}$ correspond with $\{t_-, 0, t_+\}$. This gives for B, C the values

$$(5.23)\qquad B=\frac{w_+^2t_--w_-^2t_+}{w_+w_-(w_+-w_-)}, \qquad C=\frac{t_+w_--t_-w_+}{w_-w_+(w_+-w_-)}.$$

The critical points $w_\pm$, $t_\pm$ are not analytic for small values of the parameters. However, we have the following lemma.

LEMMA 5.1. *B, C, w_+w_-, $w_0=w_++w_-$ are analytic functions of μ, ν in a fixed neighborhood of $(0,0)$. Moreover, $B=1+o(1)$, $C=o(1)$ near $(0,0)$.*

Proof. We use the notation of Theorem 5.1. We have $w_0=\ln(1+\mu)$ and the product w_+w_- is an even function of W. So the singularity in $W=\sqrt{q}\sqrt{\nu-\nu_1}$ is removed when we expand w_+w_- in powers of W. Using (5.2), we can write

$$(5.24)\qquad 2C=-\frac{\mu}{w_+w_-}\left[1-\frac{T\ln(1+\mu)/\mu}{w_+-w_-}\right].$$

We introduce a parameter η by writing

$$\beta^2=\nu[E+(\nu-\nu_0)\eta], \qquad E=-\frac{\mu^2}{4\nu_0}=-\frac{\nu_1}{4}.$$

Then we have $T=2\sqrt{q}\sqrt{E+\nu\eta}$; $\eta=\eta(\nu,\mu)$ is analytic in a neighborhood of $(0,0)$. Next we use $w_+-w_-=\ln[(1+\sqrt{q}Z)/(1-\sqrt{q}Z)]$. Since the factor $\sqrt{q}$ can be removed, we infer that the fraction $T/(w_+-w_-)$ is regular. It is easily verified that the expression

between square brackets in (5.24) vanishes when $\nu \to 0$ and that $w_+ w_- = -\nu F$, where $F = F(\nu, \mu)$ is analytic at $(0, 0)$, with $F = 1 + \mathcal{O}(\nu + \mu)$, as $\nu, \mu \to 0$. This proves that C is analytic at $(0, 0)$; the factor μ in the first fraction of (5.24) takes care of the vanishing of C at $(0, 0)$. A more detailed analysis shows that $C \sim -\mu/24$, $\mu \to 0$, $\nu = 0$. The proof for B now follows from the representation $B + w_0 C = (t_+ - t_-)/(w_+ - w_-) = T/(w_+ - w_-)$. At $\nu = 0$ this expression reduces to $\mu/\ln(\mu + 1) = 1 + \mathcal{O}(\mu)$, as $\mu \to 0$. □

COROLLARY 5.1. *Let ψ of* (5.20) *be analytic in a domain containing the points $w_\pm$. Then the coefficients a_k, b_k are analytic functions of the parameters μ, ν.*

Proof. This follows from the fact that sum and product of $w_\pm$ occur in $V(w)$ and that the Cauchy-type integrals in (5.21) are analytic functions of $w_+ + w_-$ and $w_+ w_-$. □

After these preparations we are ready to consider the following theorem.

THEOREM 5.2. *The function $t(w, \nu, \mu)$ defined by* (5.12), *with β^2 defined in* (5.13), *is analytic in a fixed neighborhood of* $(0, 0, 0)$.

Proof. We write (5.12) in the form

$$F(t, w, \mu, \nu) = tH(w) - S(t) = 0, \tag{5.25}$$

where

$$H(w) = -\mu \ln \frac{e^w - 1}{w} + (\mu + 1)w + \frac{\nu}{e^w - 1} - A, \qquad S(t) = t^2 + \beta^2 - \mu t \ln \frac{t}{w}. \tag{5.26}$$

Using (5.22) we can consider F as a function of w, with two known parameters μ, ν, and one unknown parameter y. We expand F as in (5.20):

$$F = \sum_{k=0}^{\infty} [u_k + w v_k] V^k(w), \tag{5.27}$$

where the coefficients u_k, v_k do not depend on w and t; they do depend on y, however. The first coefficients are

$$\begin{aligned}
u_0 &= -C^2 b^2 a - C^2 a^2 - aB^2 - 2aBCb + f_0 B + g_0 aC - \beta^2 \\
&\quad + \mu a(Cc_0 + Bd_0 + Cbd_0), \\
v_0 &= -2aBC - C^2 b^3 - bB^2 + f_0 C + g_0 B - 2C^2 ab - 2b^2 BC + g_0 bC \\
&\quad + \mu(Cad_0 + Bc_0 + Cbc_0 + bBd_0 + Cb^2 d_0), \\
u_1 &= -B^2 - C^2 b^2 - 2aBy - 2Cbay - 2C^2 a + f_0 y + f_1 B - 2BCb + g_0 C + g_1 aC \\
&\quad + \mu(aBd_1 + Cbad_1 + Cac_1 + Cc_0 + Bd_0 + Cbd_0 + ayd_0), \\
v_1 &= -2bBy - 2Cb^2 y - 2BC - 2Cay - 2C^2 b + g_0 y + f_1 C + g_1 B + g_1 bC \\
&\quad + \mu(bBd_1 + Cb^2 d_1 + Bc_1 + Cad_1 + Cbc_1 + Cd_0 + yc_0 + byd_0), \\
u_2 &= -ay^2 + g_1 C + f_2 B + f_1 y - 2By + g_2 aC - C^2 - 2Cby \\
&\quad + \mu(Cac_2 + aBd_2 + Bd_1 + Cc_1 + yd_0 + Cbd_1 + Cbad_2 + ayd_1), \\
v_2 &= g_2 bC - by^2 + g_1 y - 2Cy + g_2 B + f_2 C \\
&\quad + \mu(Cb^2 d_2 + byd_1 + Cbc_2 + Cad_2 + Bc_2 + bBd_2 + yc_1 + Cd_1),
\end{aligned}$$

where a, b are defined by $w^2 = a + bw + V(w)$, i.e., $a = -w_+ w_-$, $b = w_+ + w_-$ and the coefficients c_k, d_k, f_k, g_k occur in the expansions

$$wH(w) = f_0 + g_0 w + f_1 V + g_1 wV + f_2 V^2 + g_2 wV^2 + \cdots,$$

$$\ln \frac{t}{w} = c_0 + d_0 w + c_1 V + d_1 wV + c_2 V^2 + d_2 wV^2 + \cdots.$$

The coefficients u_0, v_0 vanish identically. This can be verified by straightforward manipulations. It also follows from the observation that the representation (5.22) can be viewed as a truncated expansion for t, in which the first coefficients B, C are defined properly. If more coefficients D, E, $\cdots$ had been included in y (and defined properly), more and more coefficients u_k, v_k would vanish identically. When using (5.22), only a few coefficients will vanish. Although u_1, v_1 contain the parameter y (also via c_1, d_1), these coefficients vanish too. Again, this can be verified by straightforward manipulations.

It follows that we can proceed with the equation $G=0$, where

$$G=G(y, w, \mu, \nu)=\frac{F(t, w, \mu, \nu)}{V^2(w)}.$$

The coefficient u_2 contains a term $-2By$, with B given in (5.23). From Lemma 5.1, it follows that B is bounded away from zero when the parameters μ, ν are small. The remaining contributions to u_2 containing the parameter y tend to zero as μ, $\nu \to 0$. All coefficients u_k, v_k are analytic functions of μ, ν, and the convergent infinite series (including coefficients v_2 and higher) represents a function of y, w, μ, ν that is analytic in a neighborhood of $(0, 0, 0, 0)$. Consequently, since $\partial G(0, 0, 0, 0)/\partial y = -2$, we can solve for y and this solution is an analytic function of w, μ, ν in a fixed neighborhood of $(0, 0, 0)$. The same holds for t given in (5.22). □

Remark 5.3. A simpler version ($\mu = 0$) of the above theorem is considered in Theorem 4.1. Another simpler version ($\nu = 0$) is given by [5, Thm. 2.1].

We still have to show that Ω (the image of strip H of (4.13) under the mapping $w \mapsto t$ defined in (5.12)) is large enough to contain a disc around t_+ with radius $\rho(1+t_+)^\kappa$, $\kappa \geqq \frac{1}{2}$, ρ fixed. It is not difficult to verify that when $\beta > \mu$ the proof runs as in § 4.2. If μ is much larger than β, the situation improves, and we can take $\kappa = 1$.

We conclude by computing a bound for the quantity $M_r(\beta, \mu)$ used in (5.8), and defined as in (3.16). The t-values on the circle C_r are written as $t = t_+ + \tau\sqrt{t_+ + 1}$, with $|\tau| = r$, r fixed. We assume that at least one of the parameters ν, μ is large. We have

$$t+\frac{\beta^2}{t}-\mu \ln t \sim t_+ + \frac{\beta^2}{t_+}-\mu \ln t_+ + \frac{T(1+t_+)}{t_+(\mu+T)}\tau^2 + \mathcal{O}(t_+^{-1}).$$

We denote the factor multiplying τ^2 by q. Observe that, roughly speaking, q belongs to the interval $[\frac{1}{2}, 1]$. Using this in (5.12), we obtain

$$q\tau^2 \sim -\mu \ln \frac{e^w-1}{e^{w_+}-1} + (\mu+1)(w-w_+) + \nu\left[\frac{1}{e^w-1}-\frac{1}{e^{w_+}-1}\right].$$

Denoting the right-hand side by $\psi(w)$, we see that $\psi(w_+) = \psi'(w_+) = 0$. A few computations give

$$\psi''(w_+) = 1 + \frac{2(\nu+W)}{(\nu+\mu+w)^2} = 1 + o(1).$$

To solve the equation $\psi(w) = q\tau^2$ we expand $\psi(w_+ + v) = \frac{1}{2}v^2\psi''(w_+) + \cdots$. We can take the fixed number r as small as we please. Then the solution of the above equation reads $w \sim w_+ + \tau\sqrt{2q}$. Using this in (5.16), we infer that $f(t) \sim 4\sqrt{t_+}/\tau$, under the condition that $t \in C_r$ and that at least one of the parameters ν, μ is large. Consequently, we can find a fixed number K, such that

$$M_r(\beta, \mu) \leqq K\sqrt{1+t_+}/r, \qquad \nu, \mu \in [0, \infty).$$

REFERENCES

[1] SH-N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Springer-Verlag, Berlin, New York, 1982.
[2] C. L. FRENZEN AND R. WONG, *Uniform asymptotic expansions of Laguerre polynomials*, SIAM J. Math. Anal., 19 (1988), pp. 1232-1248.
[3] A. I. MARKUSHEVICH, *Theory of Functions of a Complex Variable*, Prentice-Hall, Englewood Cliffs, NJ, 1965.
[4] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
[5] N. M. TEMME, *Laplace type integrals: transformation to standard form and uniform asymptotic expansions*, Quart. Appl. Math., XLIII (1985), pp. 103-123.
[6] ———, *Laguerre polynomials: asymptotics for large degree*, Report AM-R8610, Centre for Mathematics and Computer Science, Amsterdam, 1986.
[7] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Fourth edition, Cambridge University Press, London, New York, 1927.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 262-269, January 1990
© 1990 Society for Industrial and Applied Mathematics
014

AN ASYMPTOTIC PROBLEM IN DERANGEMENT THEORY*

J. GILLIS†, MOURAD E. H. ISMAIL‡, AND T. OFFER†

Abstract. N elements, divided into sets of respective cardinalities $\{n_1, n_2, \cdots, n_a\}$, k sets of each size, where $N = k\sum_{i=1}^{a} n_i$, are given. The probability is considered that a random permutation of the N elements is a derangement, i.e., that it leaves none of the elements in the set to which it belonged initially. In particular, an asymptotic estimate of this probability as $k \to \infty$ is obtained.

It is known that the number of possible derangements can be represented by an integral involving products of Laguerre polynomials. The probability is obtained by asymptotic evaluation of a more general integral, involving the generalized Laguerre polynomials $L_n^{(\alpha)}(x)$, of which the integral required here is a special case.

Key words. derangement, Laguerre polynomials, asymptotic

AMS(MOS) subject classifications. primary 05A05; secondary 26C05

1. Introduction. Given finite sets, of respective cardinalities $\{n_1, n_2, \cdots, n_a\}$, k sets of each size, we consider permutations of the entire set of $k\sum_{i=1}^{a} n_i$ elements. A permutation will be called a derangement if none of the elements is left in the set to which it initially belonged. Our main purpose is to estimate asymptotically, as $k \to \infty$, the probability that a random permutation of the elements is a derangement.

Let $D_k(n_1, n_2, \cdots, n_a)$ denote the total number of possible derangements and let $P_k(n_1, \cdots, n_a)$ denote the probability of a permutation being a derangement. It is clear that

$$P_k(n_1, \cdots, n_a) = D_k(n_1, \cdots, n_a) \Big/ \left(k \sum_{i=1}^{a} n_i\right)!. \tag{1.1}$$

Now it is known ([1, p. 135], [2, p. 4]) that

$$D_k(n_1, \cdots, n_a) = \prod_{i=1}^{a} \{(-1)^{n_i} n_i!\}^k \int_0^\infty \left\{\prod_{i=1}^{a} L_{n_i}(x)\right\}^k e^{-x}\, dx \tag{1.2}$$

where $L_n(x)$ denotes the Laguerre polynomial

$$L_n(x) = \sum_{m=0}^{n} (-1)^m \binom{n}{m} \frac{x^m}{m!}. \tag{1.3}$$

Since neither $D_k(n_1, \cdots, n_a)$ nor $P_k(n_1, \cdots, n_a)$ have closed-form expressions, integral representations such as (1.2) are useful for several reasons. First, we can use (1.2) and the recurrence relations for Laguerre polynomials to derive recurrence relations for $D_k(n_1, \cdots, n_a)$. For a number of such recurrence relations, some of which seem difficult to prove from direct combinatorial considerations, see [2, p. 142]. Second, an integral representation such as (1.2) proves the positivity of the linearization of products of $\{(-1)^n L_n(x)\}$, and hence the existence of a discrete convolution structure associated with $\{(-1)^n L_n(x)\}$. Third, as we will see in this paper, (1.2) can be used to estimate the size of the D_k's or P_k's for large k. We write

$$s_r = \sum_{i=1}^{a} n_i^r \qquad (r = 1, 2, \cdots), \tag{1.4}$$

* Received by the editors November 2, 1987; accepted for publication (in revised form) March 7, 1989.
† Weizmann Institute of Science, Rehovot, Israel.
‡ University of South Florida, Tampa, Florida 33620. This author's research was sponsored by the National Science Foundation.

and so

$$P_k(n_1, \cdots, n_a) = (-1)^{ks_1}\left(\prod_{i=1}^{a} n_i!\right)^k \{(s_1k)!\}^{-1} \int_0^\infty \left\{\prod_{i=1}^{a} L_{n_i}(x)\right\}^k e^{-x}\, dx. \tag{1.5}$$

We will show that, as $k \to \infty$,

$$P_k(n_1, \cdots, n_a) = \exp\left(-\frac{s_2}{s_1}\right)\left\{1 - \frac{s_1(2s_3 - s_2) - s_2^2}{2s_1^3 k} + O(k^{-2})\right\}. \tag{1.6}$$

It follows, in particular, that in the special case $a = 1$,

$$\lim_{k\to\infty} P_k(n_1) = e^{-n_1}, \tag{1.7}$$

a result previously obtained by Askey, Ismail, and Rashed [1, p. 5], though the method of proof there does not seem to extend to the case $a > 1$.

The limiting relation (1.7) is interesting because it shows that $P_k(n_1) \sim (P_k(1))^{n_1}$, i.e., the probability of having a derangement of type $n_1, n_1, \cdots, n_1$ (k times), $k \to \infty$, is asymptotically equal to that of having n_1 independent copies of the classical derangement problem with $k \to \infty$. Now (1.6), which we will prove, is more interesting because it shows that

$$\{P_k(n_1, n_2, \cdots, n_a)\}^{s_1} \sim \prod_{i=1}^{a} \{P_k(n_i)\}^{n_i},$$

a surprising result with an obvious combinatorial interpretation.

Equation (1.6) will follow from the asymptotic estimation, as $k \to \infty$, of the integral in (1.2). However, we will begin by examining a more general integral, namely,

$$I(\alpha) = \int_0^\infty \left\{\prod_{i=1}^{a} (-1)^{n_i} L_{n_i}^{(\alpha)}(x)\right\}^k x^\alpha e^{-x}\, dx \tag{1.8}$$

where the $L_{n_i}^{(\alpha)}$ are generalized Laguerre polynomials defined (for $\alpha > -1$) by

$$L_n^{(\alpha)}(x) = \sum_{m=0}^{n} (-1)^m \binom{n+\alpha}{n-m} \frac{x^m}{m!}. \tag{1.9}$$

It is not known whether the integrals $I(\alpha)$ of (1.8) are nonnegative for $\alpha > -1$. When $\alpha = 0$ the nonnegativity of $I(\alpha)$ follows from the combinatorial interpretation of [2]. The nonnegativity of $I(\alpha)$ for $\alpha \neq 0$ is difficult to prove because each Laguerre polynomial $L_n^\alpha(x)$ in the integrand has n simple zeros in the range of integration. It will be shown that, for large k,

$$I(\alpha) = (2\pi)^{1/2}(ks_1)^{ks_1+\alpha+1/2}\left\{\prod_{i=1}^{a} n_i!\right\}^{-k} \exp\left\{-ks_1 - \frac{s_2 + \alpha s_1}{s_1}\right\}\left\{1 + \frac{h}{k} + O(k^{-2})\right\} \tag{1.10}$$

where

$$h = \{(1 + 12\alpha + 18\alpha^2)s_1^2 - 12s_1s_3 + 6s_2^2 + 6(\alpha+1)s_1s_2\}/12s_1^3.$$

In particular,

$$\begin{aligned} I(0) = (2\pi)^{1/2}(ks_1)^{ks_1+1/2}&\left\{\prod_{i=1}^{a} n_i!\right\}^{-k} \exp\left\{-ks_1 - \frac{s_2}{s_1}\right\} \\ &\cdot\left\{1 - \frac{1}{12ks_1} - \frac{s_1(2s_3 - s_2) - s_2^2}{2ks_1^3} + O(k^{-2})\right\}. \end{aligned} \tag{1.11}$$

Now,

$$P_k(n_1, \cdots, n_a) = \left\{ \prod_{i=1}^{a} n_i ! \right\}^k \{(ks_1)!\}^{-1} I(0) \quad \text{(by (1.5))}$$

$$= (2\pi)^{1/2} (ks_1)^{ks_1+1/2} \{(ks_1)!\}^{-1}$$

$$\cdot \left\{ 1 - \frac{1}{12ks_1} - \frac{s_1(2s_3 - s_2) - s_2^2}{2ks_1^3} + O(k^{-2}) \right\} \tag{1.12}$$

$$\cdot \exp\left\{ -ks_1 - \frac{s_2}{s_1} \right\},$$

while, by Stirling's Formula,

$$(ks_1)! = (2\pi)^{1/2} (ks_1)^{ks_1+1/2} \exp(-ks_1) \left\{ 1 + \frac{1}{12ks_1} + O(k^{-2}) \right\}. \tag{1.13}$$

Estimate (1.6) now follows immediately from (1.12) and (1.13). It remains to establish (1.10).

2. Proof of (1.10). We will use the Laplace approach (cf. [4, p. 81]), writing

$$I(\alpha) = \int_0^\infty e^{f(x)}\, dx \tag{2.1}$$

where

$$f(x) = k \sum_{i=1}^{a} \ln\{(-1)^{n_i} L_{n_i}^{(\alpha)}(x)\} + \alpha \ln x - x. \tag{2.2}$$

It will turn out that interest centres entirely on large values of x, where $(-1)^{n_i} L_{n_i}^{(\alpha)}(x) > 0$, and we may therefore ignore questions about logarithms of negative quantities.

2.1. Maximum of the integral. Let $Q(x) = \prod_{i=1}^{a} \{(-1)^{n_i} L_{n_i}^{(\alpha)}(x)\}$. Then

$$f(x) = k \ln Q(x) + \alpha \ln x - x$$

and so the extremals of $f(x)$ are to be sought among the roots of the equation

$$(x - \alpha) Q(x) = kxQ'(x). \tag{2.3}$$

This is a polynomial equation of degree $s_1 + 1$. It is clear that, for large k, there will be a root close to $x = 0$, and $s_1 - 1$ roots close to those of $Q'(x)$, all bounded independently of k. On the other hand, if

$$Q(x) = a_0 x^{s_1} + a_1 x^{s_1-1} + \cdots,$$

then (2.3) becomes

$$kx(s_1 a_0 x^{s_1-1} + \cdots) = (x - \alpha)(a_0 x^{s_1} + a_1 x^{s_1-1} + \cdots),$$

i.e.,

$$a_0 x^{s_1+1} - \{(ks_1 + \alpha) a_0 - a_1\} x^{s_1} + \cdots = 0$$

so that the sum of all $(s_1 + 1)$ roots is $ks_1 + \alpha - a_1/a_0$. It follows that the remaining root will be

$$x_0 = ks_1 + O(1). \tag{2.4}$$

For large k this will clearly be the largest root and it is easily verified that $f(x)$ will attain its global maximum there. It remains to estimate x_0 more precisely.

For any fixed $\alpha > -1$, write

$$v_n(x) = L_n^{(\alpha)\prime}(x)/L_n^{(\alpha)}(x); \tag{2.5}$$

then (2.3) becomes

$$k \sum_{i=1}^{a} v_{n_i}(x) = 1 - \frac{\alpha}{x}. \tag{2.6}$$

But $y = L_n^{(\alpha)}(x)$ satisfies the differential equation ([3, p. 781])

$$xy'' + (\alpha + 1 - x)y' + ny = 0 \tag{2.7}$$

and hence v_n satisfies

$$x(v_n' + v_n^2) + (\alpha + 1 - x)v_n + n = 0. \tag{2.8}$$

However, $L_n^{(\alpha)}(x)$ is a polynomial of degree n, and so v_n must be of the form

$$\frac{n}{x} + O\left(\frac{1}{x^2}\right)$$

for large x. We deduce, by successive approximation in (2.8), that

$$v_n(x) = \frac{n}{x} + \frac{n(n+\alpha)}{x^2} + \frac{n(n+\alpha)(2n+\alpha-1)}{x^3} + O(x^{-4}). \tag{2.9}$$

It follows that

$$V(x) = \sum_{i=1}^{a} v_{n_i}(x) = \frac{s_1}{x} + \frac{s_2 + \alpha s_1}{x^2} + \frac{c}{x^3} + O(x^{-4}) \tag{2.10}$$

where

$$c = 2s_3 + (3\alpha - 1)s_2 + \alpha(\alpha - 1)s_1. \tag{2.11}$$

Equation (2.6) now becomes

$$1 - \frac{\alpha}{x} = k\left(\frac{s_1}{x} + \frac{s_2 + \alpha s_1}{x^2} + \frac{c}{x^3} \cdots\right). \tag{2.12}$$

Starting from (2.4) we obtain, by successive approximation in (2.12),

$$x_0 = ks_1 + \frac{s_2 + 2\alpha s_1}{s_1} + \frac{t}{ks_1^3} + O(k^{-2}) \tag{2.13}$$

where

$$t = 2s_1s_3 - s_2^2 - s_1s_2 - \alpha(\alpha + 1)s_1^2. \tag{2.14}$$

2.2. Taylor expansion of $f(x)$ about $x = x_0$. Since

$$f'(x) = kV(x) + \frac{\alpha}{x} - 1 \tag{2.15}$$

it follows that, for $r \geqq 2$,

$$f^{(r)}(x_0) = kV^{(r-1)}(x_0) + (-1)^{r-1}(r-1)!\,\alpha x_0^{-r}. \tag{2.16}$$

Substituting in (2.16) from (2.10) and (2.13) we get, after some manipulation,

$$f(x)=f(x_0)-A(x-x_0)^2+B(x-x_0)^3-C(x-x_0)^4\cdots \tag{2.17}$$

where

$$\begin{aligned} A&=-\frac{1}{2}f''(x_0)=\frac{1}{2ks_1}\left\{1-\frac{\alpha}{ks_1}+\frac{t}{k^2s_1^4}+O(k^{-3})\right\},\\ B&=\frac{1}{6}f'''(x_0)=\frac{1}{3k^2s_1^2}\left\{1-\frac{2\alpha}{ks_1}+\frac{3t}{k^2s_1^4}+O(k^{-3})\right\}, \qquad (2.18)\\ C&=-\frac{1}{24}f^{(iv)}(x_0)=\frac{1}{4k^3s_1^3}\left\{1-\frac{3\alpha}{ks_1}+\frac{6t}{k^2s_1^4}+O(k^{-3})\right\}. \end{aligned}$$

2.3. Estimation of $e^{f(x_0)}$. By (2.2),

$$e^{f(x_0)}=x_0^{\alpha}\,e^{-x_0}\left\{\prod_{i=1}^{a}(-1)^{n_i}L_{n_i}^{(\alpha)}(x_0)\right\}^k. \tag{2.19}$$

We write

$$(-1)^nL_n^{(\alpha)}(x)=\frac{x^n}{n!}G_n(x) \tag{2.20}$$

where

$$G_n(x)=1-\frac{n(n+\alpha)}{x}+\frac{n(n-1)(n+\alpha)(n+\alpha-1)}{2x^2}+O(x^{-3})$$

so that

$$\ln G_n(x_0)=-\frac{n(n+\alpha)}{x_0}-\frac{n(n+\alpha)(2n+\alpha-1)}{2x_0^2}+O(x_0^{-3}), \tag{2.21}$$

and hence

$$\begin{aligned} \prod_{i=1}^{a}(-1)^{n_i}L_{n_i}^{(\alpha)}&=x_0^{s}\left\{\prod_{i=1}^{a}n_i!\right\}^{-1}\\ &\quad\cdot\exp\left\{-\frac{s_2+\alpha s_1}{x_0}-\frac{2s_3+(3\alpha-1)s_2+\alpha(\alpha-1)s_1}{2x_0^2}+O(x_0^{-3})\right\},\\ e^{f(x_0)}&=x_0^{\alpha}\,e^{-x_0}\left\{\prod_{i=1}^{a}(-1)^{n_i}L_{n_i}(x)\right\}^k\\ &=x_0^{ks_1+\alpha}\left\{\prod_{i=1}^{a}n_i!\right\}^{-k} \qquad (2.22)\\ &\quad\cdot\exp\left\{-x_0-\frac{k(s_2+\alpha s_1)}{x_0}-\frac{k[2s_3+(3\alpha-1)s_2+\alpha(\alpha-1)s_1]}{2x_0^2}+O(kx_0^{-3})\right\}. \end{aligned}$$

Substituting from (2.13) we get, after some manipulation,

$$e^{f(x_0)}=x_0^{ks_1+\alpha}\left\{\prod_{i=1}^{a}n_i!\right\}^{-k}\exp\left\{-ks_1-\frac{2s_2+3\alpha s_1}{s_1}-\frac{g}{2ks_1^3}+O(k^{-2})\right\} \tag{2.23}$$

where

$$g=6s_1s_3-4s_2^2-3(\alpha+1)s_1s_2-\alpha(5\alpha+3)s_1^2.$$

But

$$x_0 = ks_1\left\{1+\frac{s_2+2\alpha s_1}{ks_1^2}+\frac{t}{k^2 s_1^4}+O(k^{-3})\right\}$$

and so

$$\ln\left(\frac{x_0}{ks_1}\right)=\frac{s_2+2\alpha s_1}{ks_1^2}+\frac{2t-(s_2+2\alpha s_1)^2}{2k^2 s_1^4}+O(k^{-3}) \tag{2.24}$$

whence

$$(ks_1+\alpha)\ln\left(\frac{x_0}{ks_1}\right)=\frac{s_2+2\alpha s_1}{s_1}+\frac{p}{2ks_1^3}+O(k^{-2}) \tag{2.25}$$

where

$$p = 4s_1s_3-3s_2^2-2(\alpha+1)s_1s_2-2\alpha(\alpha+1)s_1^2, \tag{2.26}$$

$$\left(\frac{x_0}{ks_1}\right)^{ks_1+\alpha}=\exp\left\{\frac{s_2+2\alpha s_1}{s_1}+\frac{p}{2ks_1^3}+O(k^{-2})\right\}. \tag{2.27}$$

From (2.23) and (2.27) we get

$$e^{f(x_0)}=(ks_1)^{ks_1+\alpha}\left\{\prod_{i=1}^{a} n_i!\right\}^{-k}\exp\left\{-ks_1-\frac{s_2+\alpha s_1}{s_1}-\frac{q}{2ks_1^3}+O(k^{-2})\right\} \tag{2.28}$$

where

$$q = 2s_1s_3-s_2^2-(\alpha+1)s_1s_2-\alpha(3\alpha+1)s_1^2. \tag{2.29}$$

2.4. Estimation of $\int_0^\infty e^{f(x)-f(x_0)}\,dx$. It follows from (2.17) that the integral can be estimated by

$$\begin{aligned}\int_{-x_0}^{\infty} e^{-Au^2+Bu^3-Cu^4}\,du &= \left(\int_{-\infty}^{\infty}-\int_{-\infty}^{-x_0}\right)e^{-Au^2+Bu^3-Cu^4}\,du \\ &= X-Y \quad \text{(say)}.\end{aligned} \tag{2.30}$$

We begin by estimating Y:

$$\begin{aligned}0<Y&=\int_{-\infty}^{-x_0} e^{-Au^2+Bu^3-Cu^4}\,du \\ &<\int_{-\infty}^{-x_0} e^{-Au^2}\,du \quad \text{since } A, B, C, \text{ are all positive} \\ &=\int_{x_0}^{\infty} e^{-Au^2}\,du.\end{aligned} \tag{2.31}$$

But for large k, $x_0 > ks_1$, by (2.13), and hence

$$\begin{aligned}0<Y&<\int_{ks_1}^{\infty} e^{-Au^2}\,du \\ &= A^{-1/2}\,\mathrm{Erfc}\,(ks_1A^{1/2}) \\ &\sim (2ks_1)^{1/2}\,\mathrm{Erfc}\,(\sqrt{ks_1/2}) \quad \text{by (2.18)} \\ &\sim 2\pi^{-1/2}\exp(-ks_1/2)\end{aligned} \tag{2.32}$$

(cf. [3, p. 298].

To estimate X we write $v^2 = Au^2 - Bu^3 + Cu^4$ leading to

$$u = av + \beta v^2 + \gamma v^3 \cdots \tag{2.33}$$

where

$$\alpha = A^{-1/2}, \qquad \gamma = \tfrac{1}{8}A^{-7/2}(5B^2 - 4AC). \tag{2.34}$$

We then get

$$\begin{aligned} X &\sim \int_{-\infty}^{\infty} e^{-v^2}\, du \\ &= \int_{-\infty}^{\infty} e^{-v^2} \frac{du}{dv}\, dv \\ &= \int_{-\infty}^{\infty} e^{-v^2}(\alpha + 3\gamma v^2 + \cdots)\, dv \\ &\sim \sqrt{\pi}(\alpha + 3\gamma/2) \\ &= \frac{1}{16}\sqrt{\pi}\, A^{-7/2}(16A^3 + 15B^2 - 12AC). \end{aligned} \tag{2.35}$$

Substituting again from (2.18), we obtain

$$\int_0^{\infty} \exp\{f(x) - f(x_0)\}\, dx = \sqrt{2\pi k s_1}\left\{1 + \frac{1+6\alpha}{12ks_1} + O(k^{-2})\right\}. \tag{2.36}$$

2.5. Proof of (1.10) and (1.6). It follows from (2.1) that

$$\begin{aligned} I(\alpha) &= e^{f(x_0)} \int_0^{\infty} e^{f(x)-f(x_0)}\, dx \\ &= (ks_1)^{ks_1+\alpha}\left\{\prod_{i=1}^{a} n_i!\right\}^{-k} \exp\left\{-ks_1 - \frac{s_2+\alpha s_1}{s_1} - \frac{q}{2ks_1^3} + O(k^{-2})\right\} \\ &\quad \cdot \sqrt{2\pi k s_1}\left\{1 + \frac{1+6\alpha}{12ks_1} + O(k^{-2})\right\} \quad \text{by (2.28) and (2.36)} \\ &= (2\pi)^{1/2}(ks_1)^{ks_1+\alpha+1/2}\left\{\prod_{i=1}^{a} n_i!\right\}^{-k} \exp\left\{-ks_1 - \frac{s_2+\alpha s_1}{s_1}\right\} \\ &\quad \cdot \left\{1 + \frac{(1+6\alpha)s_1^2 - 6q}{12ks_1^3} + O(k^{-2})\right\} \\ &= (2\pi)^{1/2}(ks_1)^{ks_1+\alpha+1/2}\left\{\prod_{i=1}^{a} n_i!\right\}^{-k} \exp\left\{-ks_1 - \frac{s_2+\alpha s_1}{s_1}\right\}\left\{1 + \frac{h}{k} + O(k^{-2})\right\}, \end{aligned}$$

i.e., (1.10).

Equation (1.6) now follows as shown in § 1, (1.12), and (1.13).

3. Some numerical results. Table 1 may give some idea as to the accuracy of approximation (1.6) for even quite small values of k. All results are correct to the fourth decimal place.

TABLE 1

$(n_1, \cdots, n_a)$	k	Direct evaluation by (1.5)	Asymptotic approximation
2	4	0.1179	0.1184
2	6	0.1243	0.1241
3	5	0.0398	0.0398
1, 2	3	0.1703	0.1726
1, 2, 3	2	0.0779	0.0799

REFERENCES

[1] R. ASKEY, M. ISMAIL, AND T. RASHED, MRC Technical Report #1522, 1975.

[2] S. EVEN AND J. GILLIS, *Derangements and Laguerre Polynomials*, Math. Proc. Cambridge Philos. Soc., 79 (1976), pp. 135–143.

[3] M. ABRAMOWITZ AND I. A. STEGUN, EDS. *Handbook of Mathematical Functions*, Dover, New York, 1965.

[4] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 21, No. 1, pp. 270-280, January 1990
© 1990 Society for Industrial and Applied Mathematics
015

THETA FUNCTION GENERALIZATIONS OF SOME CONSTANT TERM IDENTITIES IN THE THEORY OF RANDOM MATRICES*

P. J. FORRESTER†

Abstract. The probability distribution for the location of the eigenvalues in Dyson's unitary random matrix ensembles is generalized to involve theta functions. At three special values of a parameter, corresponding to Dyson's orthogonal, unitary, and symplectic ensembles, the normalization constant of the probability distribution is calculated. The results can be viewed as giving the constant term in the Laurent expansion of the multivariable function $\prod_{k\neq l=1}^{N}(1-w_k/w_l)^{\Gamma/2}(q^2w_k/w_l;\,q^2)_\infty^\Gamma$ for $\Gamma=1$, 2, and 4.

Key words. random matrix ensembles, constant term identities, q-series

AMS(MOS) subject classifications. 26B99, 26C99

1. Introduction and summary. The probability distribution for N points on a line

$$P_{N\Gamma}(x_1, x_2, \cdots, x_N)\,dx_1\,dx_2\cdots dx_N \tag{1.1}$$

where

$$P_{N\Gamma}=C_{N\Gamma}^{-1}\left\{\prod_{1\leqq j<k\leqq N}|\theta_1(\pi(x_k-x_j)/L,\,q)|\right\}^\Gamma \tag{1.2}$$

is a natural generalization of a distribution first studied by Dyson [5]. In (1.2)

$$\begin{aligned}\theta_1(z,q)&=-i\sum_{n=-\infty}^{\infty}(-1)^n q^{(n+1/2)^2}e^{2i(n+1/2)z}\\&=2q^{1/4}\sin z\prod_{n=1}^{\infty}(1-q^{2n}e^{2iz})(1-q^{2n}e^{-2iz})(1-q^{2n})\end{aligned} \tag{1.3}$$

and $C_{N\Gamma}$ denotes the normalization constant. Dyson's probability distribution is reclaimed in the $q\to 0$ limit of (1.2), this representing the eigenvalue distribution function of three ensembles of unitary random matrices (with $\pi x/L$ identified as the phase θ) at the special couplings $\Gamma=1$, 2, and 4. These three ensembles are directly related to the three classical groups: orthogonal, unitary, and symplectic [6].

For general $q<1$, (1.2) has three further physical interpretations. First, $P_{N\Gamma}$ represents (up to a constant) the Boltzmann factor of the classical one-component plasma with the logarithmic potential, interacting on a line with doubly periodic boundary conditions. To see this, we observe that the function

$$\Phi(x,y)=\frac{1}{2\pi LW}\sum_{\substack{m,n=-\infty\\(m,n)\neq(0,0)}}^{\infty}\frac{e^{2\pi i(mx/L+ny/W)}}{(m/L)^2+(n/W)^2} \tag{1.4}$$

satisfies the two-dimensional Poisson equation

$$\nabla^2\Phi(x,y)=-2\pi\delta(\mathbf{x}) \tag{1.5}$$

and is periodic in x and y (periods L and W), respectively. It has been shown by Glasser [9] that (1.4) can be summed to give

$$\Phi(x,y)=\frac{\pi y^2}{(LW)}-\log\left|\theta_1\left(\pi\left(\frac{x}{L}+\frac{iy}{L}\right),\,e^{-\pi W/L}\right)\right|+\text{const.} \tag{1.6}$$

* Received by the editors July 25, 1988; accepted for publication (in revised form) March 1, 1989.
† Department of Mathematics, La Trobe University, Bundoora, Victoria 3083, Australia.

Thus with

$$q = e^{-\pi W/L}, \quad \Gamma = e^2/k_B T, \quad \text{and } y \text{ fixed} \tag{1.7}$$

where e denotes the magnitude of the charges, the plasma interpretation of (1.2) follows immediately.

The second physical interpretation of (1.2) is as the ground state wavefunction of a quantum many-body problem. This observation is due to Sutherland [14] who showed that $P_{N\Gamma}$ satisfies the Schrödinger equation

$$P_{N\Gamma}^{-1} \sum_{j=1}^{N} \frac{\partial^2 P_{N\Gamma}}{\partial (x_j)^2} = V - E_0 \tag{1.8}$$

where

$$\begin{aligned} V - E_0 = & \sum_{1 \leqq j < k \leqq N} (\Gamma^2 N(\phi^2(x_k - x_j) + \phi'(x_k - x_j)) - 2\Gamma(\Gamma - 1)\phi'(x_k - x_j)) \\ & + \Gamma^2 N(N-1)(N-2)\eta/L. \end{aligned} \tag{1.9}$$

In (1.9)

$$\phi(x) = \frac{\pi \theta_1'(\pi x/L, q)}{L\theta_1(\pi x/L, q)} \tag{1.10}$$

and

$$\eta = -\frac{\pi^2 L}{6} \frac{\theta_1'''(0)}{\theta_1'(0)}. \tag{1.11}$$

In the large L (thermodynamic) limit, the potential V can be written as two parts: the one-dimensional Coulomb potential (with periodic boundary conditions, period L)

$$V_0(x) = \Gamma^2 \left(-\frac{4N}{LW^2}|x| + \frac{4Nx^2}{L^2 W^2} \right) \tag{1.12}$$

and a short-range potential

$$V_1(x) = \Gamma^2 \frac{4Nx}{LW^2} \left(\frac{|x|}{x} - \coth\left(\frac{x}{W} \right) \right) + \frac{2\Gamma(\Gamma - 1)}{W^2} \frac{1}{\sinh^2 (x/W)}. \tag{1.13}$$

(Note that our πW is denoted r in [14].)

The third physical interpretation of (1.2) follows from the observation that when multiplied by a suitable function of the nome q, P_{N1} satisfies the N-dimensional heat equation (see (2.18) below). Further, suppose $q = e^{-4\pi^2 Dt/L^2}$, D denoting the diffusion constant. Then the function $g_N(q)\psi$ in (2.18) represents the probability that N walkers undergoing Brownian motion, initially equally spaced on a circle of circumference length L, arrive at the points $x_1, \cdots, x_N$ in time t without their paths intersecting one another. Walkers whose paths cannot intersect have been termed vicious in [7].

In this paper we address the problem of evaluating the normalization $C_{N\Gamma}$ in (1.2), which is given by the N-dimensional integration

$$C_{N\Gamma} = \left(\prod_{l=1}^{N} \int_0^L dx_l \right) \left\{ \prod_{1 \leqq j < k \leqq N} |\theta_1(\pi(x_k - x_j)/L; q)| \right\}^{\Gamma}. \tag{1.14}$$

An interesting alternative representation can be obtained by using the product form of θ_1 given in (1.3), and noting

$$|2 \sin (a-b)|^2 = (1 - e^{2i(a-b)})(1 - e^{-2i(a-b)}). \tag{1.15}$$

The integrand now consists of simple factors that can all be expanded in a multivariable Laurent series in terms of the $e^{2\pi i x_j/L} j = 1, \cdots, N$. Clearly, the only term that would give a nonzero contribution to the integral is the constant term (CT), which is independent of all the $e^{2\pi i x_j/L}$ (but of course still dependent on q). Writing

$$w_j = e^{2\pi i x_j/L} \tag{1.16}$$

we thus have

$$C_{N\Gamma} = L^N (q^2; q^2)_\infty^{\Gamma N(N-1)/2} q^{\Gamma N(N-1)/8} K_{N\Gamma} \tag{1.17}$$

where

$$K_{N\Gamma} = \mathrm{CT} \prod_{\substack{k,l=1 \\ k \neq l}}^{N} \left(1 - \frac{w_k}{w_l}\right)^{\Gamma/2} \left(q^2 \frac{w_k}{w_l}; q^2\right)_\infty^{\Gamma}, \tag{1.18}$$

the symbol CT denoting the constant term in the Laurent expansion (q is regarded as a constant). In (1.17) and (1.18) we have introduced the notation

$$(z; q)_\infty = \prod_{n=0}^{\infty} (1 - zq^n). \tag{1.19}$$

The Laurent expansion in (1.18) is to be constructed by applying the binomial theorem to each individual factor. In general, this will give a formal series, since convergence will only occur when $|w_k| = 1$, $k = 1, 2, \cdots, N$. The cases $\Gamma = 2m$, m a positive integer, are the only exceptions; the Laurent expansion is then convergent for all $w_k \neq 0$.

If $q = 0$, the right-hand side of (1.18) reduces to

$$\mathrm{CT} \prod_{\substack{k,l=1 \\ k \neq l}}^{N} \left(1 - \frac{w_k}{w_l}\right)^{\Gamma/2}, \tag{1.20}$$

which was first considered by Dyson [5]. On the basis of three exact evaluations ($\Gamma = 1$, 2, and 4), (1.20) was conjectured to equal

$$\frac{(\Gamma N/2)!}{(\Gamma/2)!^N}. \tag{1.21}$$

This was subsequently proved by Wilson [17].

Although the identity equating (1.20) and (1.21) has been extensively generalized in a number of directions ([1]-[4], [12]; [10] gives a comprehensive account of pre-1984 literature), none of these works give the value of (1.18). Below we will provide two new constant term identities that give the evaluation of (1.18) for $\Gamma = 1$, 2 and 4.

We find

$$K_{N1} = \frac{N!}{(q^2; q^2)^{(N-1)}} \begin{cases} \dfrac{1}{\pi^{(N-1)/2}} \displaystyle\prod_{j=1}^{(N-1)/2} \sum_{n=-\infty}^{\infty} \frac{q^{2Nn^2+4nj}}{Nn+j}, & N \text{ odd}, \\ \dfrac{1}{\pi^{N/2}} \displaystyle\prod_{j=1}^{N/2} \sum_{n=-\infty}^{\infty} \frac{q^{2Nn^2+4n(j-1/2)}}{Nn+j-1/2}, & N \text{ even}, \end{cases} \tag{1.22}$$

$$K_{N2} = \frac{N!}{(q^2; q^2)_\infty^{2(N-1)}} \frac{(q^{4N}; q^{4N})_\infty^N}{(q^4; q^4)_\infty}, \tag{1.23}$$

$$K_{N4} = \frac{2^N N!}{(q^2; q^2)_\infty^{4N-1}} \prod_{j=0}^{N-1} \sum_{n=-\infty}^{\infty} \left(2nN + j + \frac{1}{2}\right) q^{4Nn^2+4n(j+1/2)}. \tag{1.24}$$

The remainder of the paper consists of three parts. First, we study an identity due to Sutherland [15], which is extended to calculate a previously unspecified proportionality constant. Second, we apply generalizations of integration techniques used in the theory of random matrices [13] to deduce the identity (3.17) from which (1.22) and (1.24) follow. Third, an identity of Macdonald [11] for the root system A_N is used to derive (1.23).

2. Some determinant identities. The following result has been derived (but not formally proved) by Sutherland [15].

THEOREM 2.1. *Let N be odd and $f_N(q)$ be some yet-to-be-determined function of the nome q. With the notation*

$$\psi(x_1, \cdots, x_N; q) = \prod_{1 \leqq k < l \leqq N} \theta_1(\pi(x_l - x_k); q) \tag{2.1}$$

we have

$$f_N(q)\psi(x_1, \cdots, x_N; q) = \int_0^1 d\gamma \det[\theta_3(\pi(x_j + \gamma - l/N); q^{1/N})]_{j,l=1,\cdots,N}. \tag{2.2}$$

The θ_1 function is defined by (1.3) *while*

$$\theta_3(z; q) = \sum_{n=-\infty}^{\infty} q^{n^2} e^{2iz}. \tag{2.3}$$

(This is the notation of Whittaker and Watson [16].)

Proof. Both the left-hand side (LHS) and right-hand side (RHS) of (2.2) are antisymmetric functions of $x_1, \cdots, x_N$ that vanish whenever $x_k = x_{k'}$, $k, k' = 1, 2, \cdots, N$ $(k \neq k')$. It thus suffices to check that both sides of (2.2) are the same function of x_1 with $x_2, \cdots, x_N$ regarded as fixed. We do this by studying the periodicity properties of each side.

From the definitions (1.3) and (2.2), and noting N is odd, it is immediately obvious that both the LHS and RHS are periodic under the translation $x_1 \mapsto x_1 + 1$. Now write $q = e^{\pi i \tau}$, where $\mathrm{Im}(\tau) > 0$, and consider the periodicity of both sides under the mapping $x_1 \mapsto x_1 + \tau$. Since

$$\theta_1(\pi(x + \tau); q) = -q^{-1} e^{-2\pi i x} \theta_1(\pi x; q) \tag{2.4}$$

the LHS remains the same apart from a factor

$$q^{-(N-1)} e^{-2\pi i x_1 (N-1)} \prod_{l=2}^{N} e^{2\pi i x_l}. \tag{2.5}$$

After deleting the minus sign prefactor, (2.4) holds with θ_1 replaced by θ_3. Thus, after replacing x_1 by $x_1 + \tau$, the lth term of the first row of the determinant in (2.2) can be written

$$q^{-(N-1)^2/N} e^{-2\pi i (x_1 + \gamma - l/N)(N-1)} \theta_3(\pi(x_1 + \gamma + \tau/N - l/N), q^{1/N}). \tag{2.6}$$

On the other hand, according to (2.4), the lth term of the jth row $(j = 2, \cdots, N)$ can be written as follows:

$$q^{1/N} e^{2\pi i (x_j + \gamma - l/N)} \theta_3(\pi(x_j + \gamma + \tau/N - l/N), q^{1/N}). \tag{2.7}$$

If in the first row we note that $e^{2\pi i l(N-1)/N} = e^{-2\pi i l/N}$, a common factor of $e^{-2\pi i l/N}$ can be removed from the lth column (the product from $l = 1, \cdots, N$ of such a factor

equals 1). Furthermore, removing obvious common factors from each row of the determinant, the RHS becomes

$$q^{-(N-1)}\, e^{-2\pi i x_1 (N-1)} \left(\prod_{l=2}^{N} e^{2\pi i x_l} \right) \cdot \int_0^1 d\gamma \, \det\left[\theta_3(\pi(x_j + \gamma + \tau/N - l/N); q^{1/N}) \right]_{j,l=1,\cdots,N}. \tag{2.8}$$

The integral is in fact the same as the RHS of (2.2) since the line integral along the path $\gamma + i \operatorname{Im}(\tau)/N$, $0 \leqq \gamma \leqq 1$, is the same as that along the unit interval $0 \leqq \gamma \leqq 1$, by Cauchy's theorem and the periodicity of the integral. Comparing (2.5) and (2.8), we see that the periodicity factors under the transformation $x_1 \mapsto x_1 + \tau$ of the LHS and RHS are the same.

Finally, consider the ratio RHS/LHS. From the above results, this is a doubly periodic function with periods π and τ. Furthermore, since the zeros of both the LHS and RHS are simple and occur at $x_1 = x_2, \cdots, x_N$ (mod π and mod τ), we have that RHS/LHS is a doubly periodic entire function, and thus by Liouville's theorem is a constant. Hence we have the result (2.2). □

The corresponding result for N even (which is not given in [13]) is Theorem 2.2.

THEOREM 2.2. *Let $g_N(q)$ be some yet to be determined function of q. Then for N even*

$$g_N(q)\psi(x_1, \cdots, x_N; q) = \int_0^1 d\gamma \, \det\left[\theta_1\left(\pi\left(x_j + \gamma - \frac{l}{N} \right); q^{1/N} \right) \right]_{j,l=1,\cdots,N} \tag{2.9}$$

where ψ is as defined in (2.1).

The proof is very similar to that above, so it will not be given here.

2.1. Calculating the proportionality constants. Let us now take up the problem of calculating $f_N(q)$ and $g_N(q)$ in (2.2) and (2.9). We have Theorem 2.3.

THEOREM 2.3. *The proportionality constants in* (2.2) *and* (2.9) *are given by*

$$f_N(q) = g_N(q) = N^{N/2} q^{-(N-1)(N-2)/24} (q^2; q^2)_\infty^{-(N-1)(N-2)/2}. \tag{2.10}$$

Proof. We will make use of an identity of Macdonald [11] (see also Andrews [1]) that states

$$\mathrm{CT}\left\{ \prod_{1 \leqq j < k \leqq N} \left(\frac{w_j}{w_k}; q \right)_\infty \left(q \frac{w_k}{w_j}; q \right)_\infty \right\} = \frac{1}{(q; q)_\infty^{N-1}}. \tag{2.11}$$

The product on the LHS of (2.11) is simply related to (2.1). Using (1.3) and the notation (1.16), we see

$$\prod_{1 \leqq j < k \leqq N} \left(\frac{w_j}{w_k}; q \right)_\infty \left(q \frac{w_k}{w_j}; q \right)_\infty = i^{N(N-1)/2} q^{-N(N-1)/8} (q^2; q^2)^{-N(N-1)/2} \cdot \left(\prod_{k=1}^{N} e^{\pi i (N-2k+1) x_k} \right) \psi(x_1, \cdots, x_N; q). \tag{2.12}$$

To proceed further we must specify the parity of N. Let us suppose N is even. Our strategy is to substitute for ψ according to (2.9) and integrate from zero to 1 for each x_k, $k = 1, \cdots, N$ (which is equivalent to calculating the constant term). Since the constant term of the LHS is given by (2.11), the only unknown will be $g_N(q)$, which will thus be specified.

Note that by expanding the determinant in (2.9) row-by-row according to the series definition of θ_1, and using the definition of a determinant as a sum over permutations, we have

$$\begin{aligned}\left(\prod_{k=1}^{N}\int_0^1 dx_k\, e^{\pi i(N-2k+1)x_k}\right)\psi(x_1,\cdots,x_N;q)\\ =g_N^{-1}(q)(-i)^N\sum_{n_1=-\infty}^{\infty}\cdots\sum_{n_N=-\infty}^{\infty}\left(\prod_{j=1}^{N}(-1)^{n_j}q^{(n_j+1/2)^2/N}\right)\\ \cdot\sum_{P=1}^{N!}\varepsilon(P)\int_0^1 d\gamma\left(\prod_{j=1}^{N}\exp\left(2\pi i(\gamma-P(j)/N)(n_j+1/2)\right)\right.\\ \left.\cdot\int_0^1 dx_j\exp\left(2\pi i x_j(n_j+N/2+1-j)\right)\right).\end{aligned}\tag{2.13}$$

In the integral over x_j the only nonzero term is when

$$n_j=j-1-N/2,\qquad j=1,2,\cdots,N\tag{2.14}$$

and so the RHS of (2.13) becomes

$$g_N^{-1}(q)q^{(N^2-1)/12}\sum_{P=1}^{N!}\varepsilon(P)\prod_{j=1}^{N}\exp\left(-2\pi iP(j)(j-(N+1)/2)/N\right).\tag{2.15}$$

The sum over permutations in (2.15) is by definition equal to

$$\det\left[\exp\left(-2\pi ik(j-(N+1)/2)/N\right)\right]_{j,k=1,\cdots,N}.\tag{2.16}$$

Multiplying this determinant by its complex conjugate gives N^N, so up to a phase of unit modulus, (2.16) is equal to $N^{N/2}$. To determine the phase, we note from van der Monde's determinant expansion that (2.16) is equal to

$$\prod_{1\leqq k<j\leqq N}\left(e^{\pi i(y_j-y_k)}-e^{-\pi i(y_j-y_k)}\right),\qquad y_j=-\frac{j}{N},\tag{2.17}$$

which says immediately that the phase of (2.16) is $(-i)^{N(N-1)/2}$ and so comparison of (2.11) and (2.15) gives the evaluation of $g_N(q)$ in (2.10). An analogous argument, using (2.2) instead of (2.9), gives the same result for $f_N(q)$. □

As an aside, we note a different proof of the following remarkable result due to Sutherland [15].

THEOREM 2.4. *With $q=e^{\pi i\tau}$ and $g_N(q)\psi$ given by* (2.10) *and* (2.1),

$$\sum_{j=1}^{N}\frac{\partial^2}{\partial(x_j)^2}\left(g_N(q)\psi\right)=4\pi iN\frac{\partial}{\partial\tau}\left(g_N(q)\psi\right)\tag{2.18}$$

and similarly for $f_N(q)\psi$. That is, the functions $g_N(q)\psi$ and $f_N(q)\psi$ satisfy the N-dimensional heat equation.

Proof. We simply note that the RHS of the identity (2.9) obeys (2.18). The partial derivatives can be performed row-by-row in the determinant, and the identity follows immediately from the fact that a single θ_1 function satisfies the one-dimensional heat equation, so that

$$\frac{\partial^2}{\partial(x_j)^2}\theta_1(\pi(x_j-x_k);q)=4\pi i\frac{\partial}{\partial\tau}\theta_1(\pi(x_j-x_k);q).\tag{2.19}$$

□

Remark. Sutherland [15] has provided a different derivation of (2.18) based on the Schrödinger equation (1.8). This approach can also be used to specify the proportionality constants $g_N(q)$ and $f_N(q)$, but this was not carried through in [15].

2.2. A confluent form of the determinant identities. Consider the identity (2.9). Let $N = N_1 + 2N_2$, where N_1 is even. Write $x_{N_1+2p-1} = y_p$, $p = 1, 2, \cdots, N_2$ and take the limit $x_{N_1+2p} \to y_p$. By first dividing both sides by

$$\prod_{p=1}^{N_2} (x_{N_1+2p} - y_p) \tag{2.20}$$

and subtracting the (N_1+2p-1)th row of the determinant from the (N_1+2p)th row, we obtain

$$\begin{aligned} &g_{N_1+2N_2}(q)(\theta_1'(0,q))^{N_1}\phi(x_1,\cdots,x_{N_1},y_1,\cdots,y_{N_2};q) \\ &= \int_0^1 d \left| \left[\begin{array}{c} \theta_1(\pi(x_j+\gamma-l/(N_1+2N_2)); q^{1/(N_1+2N_2)}) \\ \theta_1(\pi(y_\alpha+\gamma-l/(N_1+2N)); q^{1/(N_1+2N_2)}) \\ \theta_1'(\pi(y_\alpha+\gamma-l/(N_1+2N_2)); q^{1/(N_1+2N_2)}) \end{array} \right]_{\substack{j=1,\cdots,N_1 \\ \alpha=1,\cdots,N_2 \\ l=1,\cdots,N_1+2N_2}} \right| \end{aligned} \tag{2.21}$$

In (2.21) we have introduced the notation

$$\begin{aligned} &\phi(x_1,\cdots,x_{N_1},y_1,\cdots,y_{N_2};q) \\ &= \prod_{1\le j<k\le N_1} \theta_1(\pi(x_k-x_j);q) \prod_{j=1}^{N_2}\prod_{k=1}^{N_1} \theta_1^2(\pi(y_j-x_k);q) \\ &\quad \cdot \prod_{1\le j<k\le N_2} \theta_1^4(\pi(y_k-y_j);q) \end{aligned} \tag{2.22}$$

and the first entry in the determinant holds true for the first N_1 rows ($j = 1, \cdots, N_1$), while the rows $N_1+2\alpha-1$ are given by the second term, and the rows $N_1+2\alpha$ by the third term.

3. The constant term identities. We are now in a position to evaluate the following multidimensional integrals (or equivalently, calculate the following constant term identities):

$$\begin{aligned} I_1 &\equiv \left(\prod_{j=1}^{N}\int_0^1 dx_j\right)\left(\prod_{k=1}^{N_2}\int_0^1 dy_k\right)|\phi(x_1,\cdots,x_{N_1},y_1,\cdots,y_{N_2};q)| \\ &= q^{N_1(N_1-1)/8+N_1N_2/2+N_2(N_2-1)/2}(q^2;q^2)_\infty^{N_1(N_1-1)/2+2N_1N_2+2N_2(N_2-1)} X_{N_1,N_2} \end{aligned} \tag{3.1}$$

where

$$\begin{aligned} X_{N_1,N_2} = \mathrm{CT}\Bigg\{ &\prod_{\substack{k,l=1 \\ k\ne l}}^{N_1} \left(1-\frac{w_k}{w_l}\right)^{1/2}\left(q^2\frac{w_k}{w_l};q^2\right)_\infty \prod_{k=1}^{N_1}\prod_{\alpha=1}^{N_2}\left(1-\frac{w_k}{z_\alpha}\right)\left(1-\frac{z_\alpha}{w_k}\right) \\ &\cdot \left(q^2\frac{w_k}{z_\alpha};q^2\right)_\infty^2\left(q^2\frac{z_\alpha}{w_k};q^2\right)_\infty^2 \prod_{\substack{\alpha,\beta=1 \\ \alpha\ne\beta}}^{N_2}\left(1-\frac{z_\alpha}{z_\beta}\right)^2\left(q^2\frac{z_\alpha}{z_\beta};q^2\right)_\infty^4 \Bigg\} \end{aligned} \tag{3.2}$$

and

$$\begin{aligned} I_2 &\equiv \left(\prod_{j=1}^{N}\int_0^1 dx_j\right)|\psi(x_1,\cdots,x_N)|^2 \\ &= q^{N(N-1)/4}(q^2;q^2)^{N(N-1)} K_{N2} \end{aligned} \tag{3.3}$$

where, from the notation (1.18),

$$K_{N2} = \mathrm{CT} \prod_{\substack{k,l=1 \\ k\neq l}}^{N} \left(1-\frac{w_k}{w_l}\right)\left(q^2\frac{w_k}{w_l};\, q^2\right)_\infty^2. \tag{3.4}$$

In (3.1) and (3.3), ϕ and ψ are given by (2.22) and (2.1), respectively.

3.1. Further identities for ψ and ϕ. To evaluate (3.1) and (3.3) we must further transform the identities (2.9) and (2.21). The key step (again due to Sutherland [15]) is to expand the determinants row-by-row using (1.3) to obtain a formula analogous to (2.13).

In (2.9) we multiply the resulting expression by

$$i^{-N(N-1)/2}N^{-N}/2 \det\left[e^{2\pi i l(k-1/2)/N}\right]_{\substack{l=1,2,\cdots,N \\ k=-N/2,\cdots,N/2-1}}. \tag{3.5}$$

According to the discussion between (2.15) and (2.18), (3.5) is equal to unity. The multiplication gives

$$\begin{aligned} &g_N(q)\psi(x_1,\cdots,x_N;q) \\ &\quad = i^{-N(N-1)/2}N^{-N/2}q^{(N^2-1)/12} \\ &\qquad \cdot \int_0^1 d\gamma \det\left[e^{2\pi i x_j(k+1/2)}\theta_3(\pi N(x_j+\gamma)+\pi\tau(k+\tfrac{1}{2}),\, q^N)\right]_{\substack{j=1,2,\cdots,N \\ k=-N/2,\cdots,N/2-1}}. \end{aligned} \tag{3.6}$$

The kth and k'th ($k\neq k'$) member of each row j are now orthogonal on the interval [0, 1].

The procedure of multiplying by a determinant of the form (3.3) (this time of dimension N_1+2N_2), allows the following result to be deduced from (2.21):

(3.7)

$$\begin{aligned} &g_{N_1+2N_2}(q)(\theta_1'(0,q))^{N_2}\phi(x_1,\cdots,x_{N_1},y_1,\cdots,y_{N_2};q) \\ &= i^{-N_1/2}(N_1+2N_2)^{-(N_1+2N_2)/2}q^{[(N_1+2N_2)]^2-1]/12} \\ &\quad \cdot \int_0^1 d\gamma \det\begin{bmatrix} e^{2\pi i x_j(k+1/2)}\theta_3(\pi(N_1+2N_2)(x_j+\gamma)+\pi\tau(k+\frac{1}{2});\, q^{(N_1+2N_2)}) \\ e^{2\pi i y_j(k+1/2)}\theta_3(\pi(N_1+2N_2)(y_\alpha+\gamma)+\pi\tau(k+1/2);\, q^{(N_1+2N_2)}) \\ e^{2\pi i y_j(k+1/2)}\theta_3'(\pi(N_1+2N_2)(y_\alpha+\gamma)+\pi\tau(k+1/2);\, q^{(N_1+2N_2)}) \end{bmatrix}_{\substack{j=1,\cdots,N_1 \\ \alpha=1,\cdots,N_2 \\ k=N_2-N_1/2,\cdots, \\ N_2+N_1/2-1}} \end{aligned}$$

Here we have adopted the same convention of ordering the rows in the determinant as in (2.21).

3.2. Evaluation of I_1. To perform the integration in (3.1) we use an extension of the method of integration over alternate variables [13] used to compute I_1 in the $q\to 0$ limit by Forrester in [8].

The integrand in (3.1) is symmetric in $x_1,\cdots,x_N$ so the ordering

$$0\leqq x_1<x_2<\cdots<x_N\leqq 1 \tag{3.8}$$

can be made provided we multiply by $N!$. For the integrand we substitute the identity (3.7). From the structure of the determinant in (3.7), the integration over x_1 from zero to x_2 can be performed by integrating each term in the first row. Next integrate over

x_3 from x_2 to x_4 by integrating every term in the third row. Since

$$\int_{x_2}^{x_4} dx = \int_0^{x_4} dx - \int_0^{x_2} dx \tag{3.9}$$

we see that by adding the first row to the third, the integration can be taken from zero to x_4. Proceeding in this fashion until $x_1, x_3, \cdots, x_{N_1-1}$ have been integrated over gives for the kth entry of the $(2j-1)$th row $(j=1, 2, \cdots, N_1/2)$

$$\int_0^{x_{2j}} dx\, e^{2\pi i x(k+1/2)} \theta_3\left(\pi(N_1+2N_2)(x+\gamma)+\pi\tau\left(k+\frac{1}{2}\right); q^{(N_1+2N_2)}\right). \tag{3.10}$$

The integrand is now symmetric in $x_2, x_4, \cdots x_{N_1}$ so the ordering implicit in (3.8) can be removed, provided we divide by $(N_1/2)!$. Next we write the determinant as a sum over permutations, and write each of the θ-functions in series form. Ordering the permutations $P(2l) > P(2l-1)$, $l=1, 2, \cdots, N_1/2+N_2$ gives the expression

$$\begin{aligned} I_1 = {} & 2^{-N_1/2+N_2}\pi^{-N_1/2}(N_1+2N_2)^{-(N_1+2N_2)/2} q^{[(N_1+2N_2)^2-1]/12} \\ & \cdot g_{N_1+2N_2}^{-1}(q)(\theta_1'(0,q))^{-N_2} \sum_{n_1=\infty}^{\infty} \cdots \sum_{n_{N_1+2N_2}=-\infty}^{\infty} \sum_{P(2l)>P(2l-1)} \varepsilon(P) \\ & \cdot \prod_{k=1}^{N_1/2} \left(\frac{1}{(N_1+2N_2)n_{P(2k)}+P(2k)+\frac{1}{2}} - \frac{1}{(N_1+2N_2)n_{P(2k-1)}+P(2k-1)+\frac{1}{2}} \right) \\ & \cdot \prod_{j=1}^{N_2} ((N_1+2N_2)(n_{P(2k)}-n_{P(2k-1)})+P(2k)-P(2k-1)) \\ & \cdot \int_0^1 d\gamma \prod_{l=1}^{N_1/2+N_2} q^{(N_1+2N_2)(n_{P(2l)}^2+n_{P(2l-1)}^2)+2n_{P(2l-1)}(P(2l-1)+1/2)+2n_{P(2l)}(P(2l)+1/2)} \\ & \cdot \int_0^1 dx \exp(2\pi i(x+\gamma)[(N_1+2N_2)(n_{P(2l-1)}+n_{P(2l)})+P(2l)+P(2l-1)+1] \end{aligned} \tag{3.11}$$

where for each $l=1, 2, \cdots, N_1+2N_2$

$$P(l) \in \{-N_1/2-N_2, -N_1/2-N_2+1, \cdots, N_1/2+N_2-1\}. \tag{3.12}$$

From the integration over x in (3.11) and the allowed values of $P(l)$ in (3.12), we see immediately that the only nonzero terms in (3.11) occur when, for each $l=1, 2, \cdots, N_1/2+N_2$,

$$n_{P(2l-1)} = -n_{P(2l)}, \tag{3.13}$$

$$P(2l) = Q(l), \qquad P(2l-1) = -Q(l)-1 \tag{3.14}$$

where

$$Q(l) \in \{0, 1, \cdots, N_1/2+N_2-1\}. \tag{3.15}$$

Since the index on the sum over the n's is arbitrary, we are free to relabel and write $n_{P(j)} = n_j$ in all cases. Furthermore, all permutations given by (3.14) have even parity. Hence (3.11) reduces to

$$\begin{aligned} I_1 = {} & \pi^{-N_1/2}(N_1+2N_2)^{-(N_1+2N_2)/2} q^{((N_1+2N_2)^2-1)/12} g_{N_1+2N_2}^{-1}(q)(\theta_1'(0,q))^{-N_2} \\ & \cdot \sum_{Q=1}^{(N_1/2+N_2)!} \prod_{k=1}^{N_1/2} \sum_{n=-\infty}^{\infty} \frac{q^{2(N_1+2N_2)n^2+4(Q(k)+1/2)n}}{(N_1+2N_2)n+Q(k)+\frac{1}{2}} \\ & \cdot \prod_{j=1}^{N_2} \sum_{m=-\infty}^{\infty} \left[(N_1+2N_2)m + Q\left(\frac{N_1}{2}+j\right)+\frac{1}{2} \right] q^{2(N_1+2N_2)m^2+4(Q(N_1/2+j)+1/2)m}. \end{aligned} \tag{3.16}$$

We observe that the sum over permutations can be formed by the multiplication into series of a product expansion of a polynomial. Substituting (3.16) into (3.1) and using (2.10) and the product expansion of $\theta_1'(0, q)$, we have thus derived the following result.

THEOREM 3.1. *Let N_1 be even, and X_{N_1,N_2} be given by* (3.2). *Then*

$$X_{N_1,N_2} = \pi^{-N_1/2}(q^2; q^2)_\infty^{-N_1-4N_2+1}\left(\frac{N_1}{2}+N_2\right)!\Big/\left[\left(\frac{N_1}{2}\right)!N_2!\right]$$

(3.17)

$$\cdot\left\{\textit{coefficient of } \xi^{N_1/2} \textit{ in the expansion of } \prod_{k=0}^{N_1/2+N_2-1}(A(k; q)+\xi B(k; q))\right\}$$

where

$$A(k; q) = \sum_{n=-\infty}^{\infty}\left[(N_1+2N_2)n+k+\frac{1}{2}\right]q^{2(N_1+2N_2)n^2+4(k+1/2)n} \tag{3.18}$$

and

$$B(k; q) = \sum_{n=-\infty}^{\infty}\frac{q^{2(N_1+2N_2)n^2+4(k+1/2)n}}{(N_1+2N_2)n+k+\frac{1}{2}}. \tag{3.19}$$

The results (1.22) (N even) and (1.24) follow immediately from (3.17) by choosing $N_2 = 0$ and $N_1 = 0$, respectively. The result (1.22) for N odd can be derived from the analogue of (3.17) with N_1 odd (the necessary identity is the same as (3.17)-(3.19), except that the quantity $N_1/2$ in (3.17) is replaced by $(N_1-1)/2$, and in (3.18) and (3.19), $(k+\frac{1}{2})$ is replaced by k).

3.3. Evaluation of I_2. The constant term K_{N2} in (3.3) can be most expediently evaluated by use of the identity

$$\prod_{1\le j<k\le N}\left(\frac{w_k}{w_j}; q^2\right)_\infty\left(q^2\frac{w_j}{w_k}; q^2\right)_\infty$$

(3.20)

$$= \frac{1}{(q^2; q^2)_\infty^{N-1}}\sum_{Y_{\{m_j\}}}\sum_{P=1}^{N!}\varepsilon(P)\prod_{l=1}^{N}q^{Nm_l^2+(N+1-2P(l))m_l}w_l^{Nm_l+l-P(l)}$$

where

$$Y_{\{m_j\}} = \left\{(m_1, \cdots, m_N): \sum_{j=1}^{N} m_j = 0,\ m_j \in \mathbf{Z} \text{ each } j = 1, \cdots, N\right\}. \tag{3.21}$$

This identity is due to Macdonald [11] and relates to the root system A_N. In the present context (for N even), (3.20) follows immediately from (3.6). To see this, rewrite the LHS of (3.6) using the product expansion of the θ_1 function (3.1), and rewrite the RHS by expanding the determinant row-by-row using the series expansion (2.3) and integrating over γ.

From (3.3) and (3.20) we see

$$K_{N2} = \frac{1}{(q^2; q^2)^{2N-2}}\sum_{Y_{\{m_j\}}}\sum_{Y_{\{n_j\}}}\sum_{P=1}^{N!}\sum_{Q=1}^{N!}\varepsilon(P)\varepsilon(Q)\prod_{l=1}^{N}w_l^{N(m_l-n_l)-(P(l)-Q(l))}$$

(3.22)

$$\cdot\prod_{l=1}^{N}q^{N(m_l^2+n_l^2)+(N+1-P(l))m_l+(N+1-Q(l))n_l}.$$

Since

$$P(l), Q(l) \in \{1, 2, \cdots, N\} \tag{3.23}$$

the only constant terms in (3.22) occur when

$$P(l) = Q(l) \quad \text{and} \quad m_l = n_l, \quad l = 1, \cdots, N. \tag{3.24}$$

All permutations P then give the same contribution, so we can choose $P(l) = l$ provided we multiply by $N!$. Hence

$$K_{N2} = \frac{N!}{(q^2;q^2)^{2N-2}} \sum_{Y_{\{m_j\}}} \prod_{l=1}^{N} q^{2Nm_l^2+2(N+1-l)m_l}$$
$$= \frac{N!}{(q^2;q^2)^{2N-2}} \int_0^1 d\alpha \prod_{k=1}^{N} \theta_3(\pi\alpha + \pi\tau(N+1-2k); q^{2N}). \tag{3.25}$$

The second line of (3.25) follows from the first by using the series expansion (2.3).

From the product expansion of θ_3 we have

$$\prod_{k=1}^{N} \theta_3(\pi\alpha + \pi\tau(N+1-2k); q^{2N}) = \frac{(q^{4N};q^{4N})_\infty^N}{(q^4;q^4)_\infty} \theta_3(\pi\alpha; q^2). \tag{3.26}$$

Thus the integration over α can be done at once to yield the desired result (1.23).

Acknowledgment. I thank the referee for pointing out that (3.20) can be found in [11].

REFERENCES

[1] G. E. ANDREWS, *Notes on the Dyson conjecture*, SIAM J. Math. Anal., 11 (1980), pp. 787-792.
[2] R. A. ASKEY, *Some basic hypergeometric extensions of integrals of Selberg and Andrews*, SIAM J. Math. Anal., 11 (1980), pp. 938-951.
[3] D. M. BRESSOUD AND I. P. GOULDEN, *Constant term identities extending the q-Dyson theorem*, Trans. Amer. Math. Soc., 291 (1985), pp. 203-228.
[4] ———, *The generalized plasma in one dimension: evaluation of a partition function*, Comm. Math. Phys., 110 (1987), pp. 287-292.
[5] F. J. DYSON, *Statistical theory of the energy level of the complex systems* I, J. Math. Phys. 3 (1962), pp. 140-156.
[6] ———, *The threefold way: algebraic structure of symmetry groups and ensembles in quantum mechanics*, J. Math. Phys., 3 (1962), pp. 1191-1198.
[7] M. E. FISHER, *Walks, walls, wetting and melting*, J. Statist. Phys., 34 (1984), pp. 668-727.
[8] P. J. FORRESTER, *An exactly solvable two component classical Coulomb system*, J. Austral. Math. Soc. Ser. B, 26 (1984), pp. 119-128.
[9] M. L. GLASSER, *The evaluation of lattice sums. III. Phase modulated sums*, J. Math. Phys., 15 (1974), pp. 188-189.
[10] K. W. KADELL, *A proof of Andrews q-Dyson conjecture for n = 4*, Trans. Amer. Math. Soc., 290 (1984), pp. 127-144.
[11] I. G. MACDONALD, *Affine root systems and Dedekind's eta function*, Invent. Math., 15 (1972), pp. 91-143.
[12] ———, *Some conjectures for root systems*, SIAM J. Math. Anal., 13 (1982), pp. 988-1007.
[13] M. L. MEHTA, *Random Matrices*, Academic Press, New York, 1967.
[14] B. SUTHERLAND, *Exact ground-state wave function for a one-dimensional plasma*, Phys. Rev. Lett., 34 (1975), pp. 1083-1085.
[15] ———, *One dimensional plasma as an example of a Wigner solid*, Phys. Rev. Lett., 35 (1975), pp. 185-188.
[16] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Fourth edition, Cambridge University Press, Cambridge, UK, 1962.
[17] K. WILSON, *Proof of a conjecture by Dyson*, J. Math. Phys., 3 (1962), pp. 1040-1043.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 281–312, March 1990
© 1990 Society for Industrial and Applied Mathematics

THE BOUNDARY LAYER FOR THE REISSNER–MINDLIN PLATE MODEL*

DOUGLAS N. ARNOLD† AND RICHARD S. FALK‡

Abstract. The structure of the solution of the Reissner–Mindlin plate equations is investigated, emphasizing its dependence on the plate thickness. For the transverse displacement, rotation, and shear stress, asymptotic expansions in powers of the plate thickness are developed. These expansions are uniform up to the boundary for the transverse displacement, but for the other variables there is a boundary layer. Rigorous error bounds are given for the errors in the expansions in Sobolev norms. As applications, new regularity results for the solutions and new estimates for the difference between the Reissner–Mindlin solution and the solution to the biharmonic equation are derived. Boundary conditions for a clamped edge are considered for most of the paper, and the very similar case of a hard simply-supported plate is discussed briefly at the end. Other boundary conditions will be treated in a forthcoming paper.

Key words. Reissner, Mindlin, plate, boundary layer

AMS(MOS) subject classifications. 73K10, 35B25

1. Introduction.

The Reissner–Mindlin model describes the deformation of a plate subject to a transverse loading in terms of the transverse displacement of the midplane and the rotation of fibers normal to the midplane [9], [10]. This linear model, as well as its generalization to shells, is frequently used for plates and shells of small to moderate thickness. Specifically, let Ω denote the region in $\mathbb{R}^2$ occupied by the midsection of the plate and ω and ϕ the transverse displacement of Ω and the rotation of the fibers normal to Ω, respectively. The Reissner–Mindlin model for the bending of a clamped isotropic elastic plate in equilibrium determines ω and ϕ as the solution of the partial differential equations

$$-\mathbf{div}\, C\, \mathcal{E}(\phi) - \lambda\bar{t}^{-2}(\mathbf{grad}\,\omega - \phi) = 0, \tag{1.1}$$

$$-\lambda\bar{t}^{-2}\,\mathrm{div}(\mathbf{grad}\,\omega - \phi) = g, \tag{1.2}$$

in Ω and the boundary conditions

$$\phi = 0,\ \omega = 0, \tag{1.3}$$

on $\partial\Omega$. Here $g\bar{t}^3$ is the transverse load force density per unit area, $\bar{t}$ is the plate thickness, $\lambda = Ek/2(1+\nu)$ with E the Young's modulus, ν the Poisson ratio, and k the shear correction factor, $\mathcal{E}(\phi)$ is the symmetric part of the gradient of ϕ, and the fourth-order tensor C is defined by

$$C\mathcal{T} = D\left[(1-\nu)\mathcal{T} + \nu\,\mathrm{tr}(\mathcal{T})\mathcal{I}\right], \quad D = \frac{E}{12(1-\nu^2)},$$

*Received by the editors September 14, 1988; accepted for publication May 1, 1989. This research was supported by National Science Foundation grants DMS–86–01489 (DNA) and DMS–87–03354 (RSF), and was partially performed at and supported by the Institute for Mathematics and its Applications.

†Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.

‡Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903.

for any 2×2 matrix $\mathcal{T}$ ($\mathcal{I}$ denotes the 2×2 identity matrix). Note that the load has been scaled so that the solution tends to a nonzero limit as $\bar{t}$ tends to zero. The Dirichlet boundary conditions (1.3) model a plate which experiences no displacement along its lateral edge. This is commonly referred to as a clamped edge (which is the terminology we adopt here), although the terms welded or built-in are perhaps more descriptive.

The Reissner–Mindlin model is an alternative to the biharmonic model for plate bending. The biharmonic model gives the transverse displacement as the solution to the boundary value problem

$$D\,\Delta^2 \omega_0 = g \quad \text{in } \Omega, \quad \omega_0 = \partial\omega_0/\partial n = 0 \quad \text{on } \partial\Omega. \tag{1.4}$$

With our scaling of the load function, the solution ω_0 is independent of the plate thickness. By contrast, the solution of the Reissner–Mindlin model depends in a complex way on the plate thickness. It is the purpose of this paper to investigate the structure of solution in its dependence on $\bar{t}$.

We shall develop asymptotic expansions with respect to $\bar{t}$ for ω and $\boldsymbol{\phi}$ (as well as other quantities associated with the solution such as the shear stress). The expansions are of the following forms[1]

$$\omega \sim \omega_0 + \bar{t}^2\omega_2 + \bar{t}^3\omega_3 + \cdots,$$
$$\boldsymbol{\phi} \sim \boldsymbol{\phi}_0 + \bar{t}^2(\boldsymbol{\phi}_2 + \chi\boldsymbol{\Phi}_0) + \bar{t}^3(\boldsymbol{\phi}_3 + \chi\boldsymbol{\Phi}_1) + \cdots.$$

Here the functions ω_i and $\boldsymbol{\phi}_i$, the interior expansion functions, are independent of $\bar{t}$. The functions $\boldsymbol{\Phi}_i$ are boundary correctors. They depend on $\bar{t}$ only through the quantity $\rho/\bar{t}$, where ρ is the distance of a point of Ω from the boundary. More specifically,

$$\boldsymbol{\Phi}_i = \hat{\boldsymbol{\Phi}}_i(\rho/\bar{t}, \theta)$$

where θ is a coordinate which roughly gives arclength along the boundary (see § 2), and the function $\hat{\boldsymbol{\Phi}}_i(\eta, \theta)$ has the form of a polynomial with respect to η times $\exp(-\sqrt{12k}\eta)$. Thus $\boldsymbol{\Phi}_i$ represents a boundary layer function, which essentially lives in a strip of width $\bar{t}$ around the boundary. Finally, χ is a cutoff function which is independent of $\bar{t}$ and identically equal to unity in a neighborhood of $\partial\Omega$.

In §§ 3 and 6 we construct all terms of these expansions. Here we summarize the results for the principal terms. The function ω_0 is the solution to the biharmonic problem above, ω_2 solves

$$D\,\Delta^2 \omega_2 = \frac{-1}{6k(1-\nu)}\,\Delta g \quad \text{in } \Omega, \quad \omega_2 = 0, \quad \frac{\partial\omega_2}{\partial n} = \frac{-1}{6k(1-\nu)}\frac{\partial}{\partial n}\,\Delta\omega_0 \quad \text{on } \partial\Omega,$$

and ω_3 solves

$$D\,\Delta^2 \omega_3 = 0 \quad \text{in } \Omega, \quad \omega_3 = 0, \quad \frac{\partial\omega_3}{\partial n} = \frac{-1}{12\sqrt{3k^3}(1-\nu)}\frac{\partial^2}{\partial s^2}\,\Delta\omega_0 \quad \text{on } \partial\Omega.$$

[1] In order not to introduce unnecessary distractions, in this introduction we use a slightly different notation than in the following sections. The ω_i and $\boldsymbol{\phi}_i$ of this section are $\lambda^{-i/2}$ times the corresponding quantities used in the remaining sections, and $\hat{\boldsymbol{\Phi}}_{i-2}(\eta, \theta)$ here is $\lambda^{-i/2}$ times $\hat{\boldsymbol{\Phi}}_{i-2}(\sqrt{\lambda}\eta, \theta)$ used later.

For the expansion of ϕ, we have $\phi_0 = \mathbf{grad}\,\omega_0$, $\phi_2 = \mathbf{grad}\,\psi$, where

$$\Delta^2 \psi = 0 \quad \text{in } \Omega, \qquad \psi = \frac{1}{6k(1-\nu)}\Delta\omega_0, \quad \frac{\partial\psi}{\partial n} = 0 \quad \text{on } \partial\Omega,$$

and

$$\hat{\boldsymbol{\Phi}}_0(\eta,\theta) = -\frac{\exp(-\sqrt{12k}\eta)}{6k(1-\nu)}\frac{\partial}{\partial s}\Delta\omega_0(0,\theta)\boldsymbol{s},$$

where $\boldsymbol{s} = \boldsymbol{s}(\theta)$ is the unit tangent vector to $\partial\Omega$.

We prove a priori estimates for all terms of the expansions in § 4, and establish error bounds for the remainders in § 5. With these results, we may easily investigate the regularity of solutions of the Reissner–Mindlin system and their limit as $\bar{t} \to 0$. Supposing that g is sufficiently smooth, we have the following estimates, in which the constant C depends on g, Ω, and the elastic constants, but is independent of $\bar{t}$. Here $\|\cdot\|_s$ and $|\cdot|_s$ denote the norms in the Sobolev spaces $H^s(\Omega)$ and $H^s(\partial\Omega)$ (see § 2).

The transverse displacement ω is regular uniformly in $\bar{t}$, but the regularity of the rotation $\boldsymbol{\phi}$ is limited by the boundary layer:

$$\|\omega\|_s \le C, \qquad \|\boldsymbol{\phi}\|_s \le C\bar{t}^{\min(0,5/2-s)}, \qquad s \in \mathbb{R}.$$

Thus all derivatives of ω remain bounded uniformly in L^2 as $\bar{t} \to 0$, while for $\boldsymbol{\phi}$, the second derivatives remain bounded in L^2, but higher derivatives will in general blow up as $\bar{t} \to 0$.

The quantity $\boldsymbol{\zeta} := \lambda\bar{t}^{-2}(\mathbf{grad}\,\omega - \boldsymbol{\phi})$, which is related to the shear stress, is often of interest. From the above expansions we get

$$\lambda^{-1}\boldsymbol{\zeta} \sim \mathbf{grad}\,\omega_2 - \boldsymbol{\phi}_2 - \chi\boldsymbol{\Phi}_0 + t(\mathbf{grad}\,\omega_3 - \boldsymbol{\phi}_3 - \chi\boldsymbol{\Phi}_1) + \cdots,$$

so it has a stronger boundary layer. Indeed, $\boldsymbol{\zeta}$ is not uniformly bounded in H^s for $s > \frac{1}{2}$:

$$\|\boldsymbol{\zeta}\|_s \le C\bar{t}^{\min(0,1/2-s)}, \quad s \in \mathbb{R}.$$

Of course, the boundary layer does not limit the regularity of $\boldsymbol{\phi}$ or $\boldsymbol{\zeta}$ at a positive distance from $\partial\Omega$ nor does it affect the smoothness of their restrictions to $\partial\Omega$. Thus

$$\|\boldsymbol{\phi}\|_{H^s(\Omega_c)} + |\boldsymbol{\phi}|_s + \|\boldsymbol{\zeta}\|_{H^s(\Omega_c)} + |\boldsymbol{\zeta}|_s \le C, \qquad s \in \mathbb{R},$$

for any compact subdomain Ω_c of Ω.

In the limit as $\bar{t} \to 0$, each of the variables ω, $\boldsymbol{\phi}$, and $\boldsymbol{\zeta}$ tends in L^2 to the leading terms of its asymptotic expansions. The number of derivatives which converge and the rate of convergence may be determined by examining the first neglected interior and boundary terms of the expansions. We get, for each $s \in \mathbb{R}$, that

$$\begin{aligned}
&\|\omega - \omega_0\|_s \le C\bar{t}^2,\\
&\|\boldsymbol{\phi} - \boldsymbol{\phi}_0\|_s \le C\bar{t}^{\min(2,5/2-s)},\\
&\|\boldsymbol{\zeta} - \lambda(\mathbf{grad}\,\omega_2 - \boldsymbol{\phi}_2)\|_s \le C\bar{t}^{\min(1/2,1/2-s)}.
\end{aligned}$$

Note that for $\boldsymbol{\phi}$ and $\boldsymbol{\zeta}$, the rate of convergence depends on the Sobolev norm under consideration. For each of the variables, taking more terms from the expansions increases the rates of convergence. For example,

$$\|\omega - \omega_0 - \bar{t}^2\omega_2\|_s \le C\bar{t}^3, \qquad \|\boldsymbol{\phi} - \boldsymbol{\phi}_0 - \bar{t}^2(\boldsymbol{\phi}_2 + \chi\boldsymbol{\Phi}_0)\|_s \le C\bar{t}^{\min(3,7/2-s)}.$$

Taking sufficiently many terms in the expansions gives approximations of any desired algebraic order of convergence in $\bar{t}$ in any desired Sobolev space (provided g is sufficiently regular).

It is also possible to use the asymptotic expansion to derive estimates in function spaces other than H^s. Thus for example, we show at the end of § 5 that

$$\|\phi\|_{W^s_\infty} \leq C\bar{t}^{\min(0,2-s)},$$

and, in particular, that $\|\phi\|_{W^2_\infty}$ is uniformly bounded. Note that this is a better estimate than we would get applying the Sobolev Embedding Theorem directly to the estimates for ϕ in H^s. It is also easy to show that

$$\|\omega - \omega_0\|_{L^\infty} \leq C\bar{t}^2, \qquad \|\phi - \phi_0\|_{L^\infty} \leq C\bar{t}^2,$$

but ζ does not in general converge in $L^\infty(\Omega)$.

The Reissner–Mindlin model is discussed in many places (under various names), although not very much attention has been devoted to the boundary layer behavior. The existence of a boundary layer is noted in [6, Chaps. 8.9–8.10] and [11, Chap. 3.5]. Assiff and Yen [2] also note the existence of a boundary layer, and use separation of variable techniques to compute the exact solution to the equations on a circular plate with a special load. This calculation exhibits the boundary layer, and may be taken as an example of our theory. Häggblad and Bathe recently studied the boundary layer in more general situations via formal techniques and numerical experiments in [7]. They also consider the effect of corners, which is not treated here. In [6], [11], and [7], the authors emphasize a reformulation of the Reissner–Mindlin system consisting of a biharmonic equation for ω (with different right-hand side than (1.4)), and a singularly perturbed Laplacian for $\operatorname{rot}\phi$. These equations are coupled through somewhat complicated boundary conditions, however, and we have preferred not to use them. As far as we know, the explicit form of the asymptotic expansions and error bounds for them are new.

2. Notation and preliminaries. The letter C denotes a generic constant, not necessarily the same in each occurrence. We assume that Ω is a smooth, bounded, and simply-connected domain in $\mathbb{R}^2$. The $L^2(\Omega)$ and $L^2(\partial\Omega)$ inner products are denoted by $(\cdot,\cdot)$ and $\langle\cdot,\cdot\rangle$ respectively. We shall use the usual L^2-based Sobolev spaces $H^s(\Omega)$ and $H^s(\partial\Omega)$, $s \in \mathbb{R}$, with norms denoted by $\|\cdot\|_s$ and $|\cdot|_s$. The reader is referred to [8] for precise definitions of these spaces and their properties, of which we recall only a few here. For $s \geq 0$, H^{-s} may be identified with the dual of $\mathring{H}^s$, the closure of C_0^∞ in H^s. If $s \geq 0$, $n \geq i \geq 0$ are real numbers, then the interpolation inequality

$$\|g\|^n_{s+i} \leq C\|g\|^{n-i}_s \|g\|^i_{s+n} \tag{2.1}$$

holds. If $g \in L^2(\Omega)$ and $\Delta^{-1} g$ denotes the unique function in $H^2(\Omega) \cap \mathring{H}^1(\Omega)$ whose Laplacian is equal to g, then

$$C^{-1}\|\Delta^{-1} g\|_{s+2} \leq \|g\|_s \leq C\|\Delta^{-1} g\|_{s+2}, \qquad s \geq 0,$$

where the constant C may depend on s and Ω, but not on g. In other words, $g \mapsto \|\Delta^{-1} g\|_{s+2}$ defines an equivalent norm on $H^s(\Omega)$ for $s \geq 0$. This is also true for $s = -1$, but slightly different negative norms are needed to extend this shift theorem

to other negative values. We define $|||g|||_s = \|\Delta^{-1} g\|_{s+2}$ for $g \in L^2(\Omega)$ and all real s. Then $|||\cdot|||_s$ is equivalent to the ordinary Sobolev norm $\|\cdot\|_s$ for $s \geq 0$ and $s = -1$. For $s = -2$, $|||\cdot|||_s$ is equivalent to the norm in the dual space of $H^2(\Omega) \cap \mathring{H}^1(\Omega)$. The norm $|||\cdot|||$ can be identified for other values of s as well, but this is not necessary for our purposes. From (2.1) we have

$$|||g|||_{s+i}^n \leq C|||g|||_s^{n-i}|||g|||_{s+n}^i,$$

valid for all real $s \geq -2$, $n \geq i \geq 0$. We shall make frequent use of this fact to bound sums of the form $\sum_{i=0}^n t^i|||g|||_{s+i}$ by a multiple of the sum of the first and last terms.

We also require the quotient space $H^s(\Omega)/\mathbb{R}$. An element $p \in H^s(\Omega)/\mathbb{R}$ is a coset consisting of all functions in $H^s(\Omega)$ differing from a fixed function by a constant. The quotient norm is given by

$$\|p\|_{s/\mathbb{R}} = \min_{q \in p} \|q\|_s.$$

In fact, $\|p\|_{s/\mathbb{R}} = \|p_0\|_s$ where p_0 is the unique function in the coset p having mean value zero.

We use boldface type to denote 2-vector-valued functions, operators whose values are vector-valued functions, and spaces of vector-valued functions. Script type is used in a similar way for 2×2-matrix objects. Thus, for example, $\operatorname{div}\boldsymbol{\psi} \in L^2(\Omega)$ for $\boldsymbol{\psi} \in \boldsymbol{H}^1(\Omega)$, while $\mathbf{div}\,\mathcal{T} \in \boldsymbol{L}^2(\Omega)$ for $\mathcal{T} \in \mathcal{H}^1(\Omega)$. Finally, we use various standard differential operators:

$$\mathbf{grad}\, r = \begin{pmatrix} \partial r/\partial x \\ \partial r/\partial y \end{pmatrix}, \qquad \operatorname{div}\boldsymbol{\psi} = \frac{\partial \psi_1}{\partial x} + \frac{\partial \psi_2}{\partial y},$$

$$\mathbf{div}\begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix} = \begin{pmatrix} \partial t_{11}/\partial x + \partial t_{12}/\partial y \\ \partial t_{21}/\partial x + \partial t_{22}/\partial y \end{pmatrix},$$

$$\mathbf{curl}\, p = \begin{pmatrix} -\partial p/\partial y \\ \partial p/\partial x \end{pmatrix}, \qquad \operatorname{rot}\boldsymbol{\psi} = \frac{\partial \psi_1}{\partial y} - \frac{\partial \psi_2}{\partial x}.$$

Note that these differential operators annihilate constants, and consequently induce operators on the quotient space $H^s(\Omega)/\mathbb{R}$ for each s. We denote the induced operator in the same way as the original. Thus, for example, if $p \in H^1(\Omega)/\mathbb{R}$, $\mathbf{curl}\, p$ denotes the element of $\boldsymbol{L}^2$ obtained by applying the curl to any element in the coset p.

In our analysis, we rely on an equivalent formulation of the Reissner–Mindlin plate equations, suggested by Brezzi and Fortin [3]. This formulation is derived by using the Helmholtz theorem to decompose the scaled transverse shear stress vector:

$$\boldsymbol{\zeta} := \lambda \bar{t}^{\,-2}(\mathbf{grad}\,\omega - \boldsymbol{\phi}) = \mathbf{grad}\, r + \mathbf{curl}\, p, \qquad r \in \mathring{H}^1(\Omega), \quad p \in H^1(\Omega)/\mathbb{R}. \tag{2.2}$$

Setting $t^2 = \bar{t}^{\,2}/\lambda$, one finds

$$-\Delta r = g, \tag{2.3}$$

$$-\mathbf{div}\, C\, \mathcal{E}(\boldsymbol{\phi}) - \mathbf{curl}\, p = \mathbf{grad}\, r, \tag{2.4}$$

$$-\operatorname{rot}\boldsymbol{\phi} + t^2 \Delta p = 0, \tag{2.5}$$

$$-\Delta\omega = -\operatorname{div}\boldsymbol{\phi} - t^2 \Delta r, \tag{2.6}$$

with the boundary conditions

$$r = 0, \qquad \boldsymbol{\phi} = 0, \qquad \frac{\partial p}{\partial n} = 0, \qquad \omega = 0. \tag{2.7}$$

Note that r satisfies a Dirichlet problem for Poisson's equation, which decouples from the other three equations. Once r has been determined, $\boldsymbol{\phi}$ and p may be computed from (2.4) and (2.5) and their boundary conditions, and then ω is determined by a second Dirichlet problem for Poisson's equation. Thus all the difficulties of the problem have been concentrated in the system (2.4)–(2.5) for $\boldsymbol{\phi}$ and p. When $t = 0$, this system of partial differential equations is very similar to the Stokes equations. For positive t, these two equations represent a singularly perturbed Stokes-like system.

It is easy to check that this reformulation is equivalent to the usual Reissner–Mindlin formulation (cf. [3] or [1]). That is, if $(\omega, \boldsymbol{\phi}) \in H^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ solves (1.1)–(1.3) and $(r, p) \in \mathring{H}^1(\Omega) \times H^1(\Omega)/\mathbb{R}$ are defined (uniquely) by (2.2), then (2.3)–(2.7) are satisfied, and, conversely, if $(\omega, \boldsymbol{\phi}, p, r) \in H^1(\Omega) \times \boldsymbol{H}^1(\Omega) \times H^1(\Omega)/\mathbb{R} \times H^1(\Omega)$ solves (2.3)–(2.7) then (1.1)–(1.3) hold.

To describe the boundary layer for the Reissner–Mindlin plate, we shall employ the standard technique of making a change of variable in a neighborhood of the boundary. Let $(X(\theta), Y(\theta))$, $\theta \in [0, L)$, be a parametrization of $\partial\Omega$ by arclength, and let Ω_0 be a normal tubular neighborhood of $\partial\Omega$ in Ω. Then, for each point $z = (x, y) \in \Omega_0$ there is a unique nearest point $z_0 \in \partial\Omega$. Let θ denote the arclength parameter, with counterclockwise orientation, corresponding to z_0 and $\rho = |z - z_0|$ the distance from the point z to the boundary. Since Ω_0 is a tubular neighborhood of $\partial\Omega$, the correspondence $(x, y) \mapsto (\rho, \theta)$ is a diffeomorphism between Ω_0 and $(0, \rho_0) \times \mathbb{R}/L$ for some $\rho_0 > 0$. Explicitly, $x = X(\theta) - \rho Y'(\theta)$, $y = Y(\theta) + \rho X'(\theta)$. A simple computation shows that the Jacobian of the transformation from (x, y) coordinates to (ρ, θ) coordinates on Ω_0 is given by $1 - \kappa(\theta)\rho$, where κ denotes the curvature of $\partial\Omega$. With these definitions, the unit outward normal and counterclockwise unit tangent vectors are given by

$$\boldsymbol{n} = -\operatorname{\mathbf{grad}} \rho = -\operatorname{\mathbf{curl}} \theta, \quad \boldsymbol{s} = \operatorname{\mathbf{grad}} \theta = -\operatorname{\mathbf{curl}} \rho \quad \text{on } \partial\Omega.$$

We use tildes to denote the corresponding change of variables for functions, i.e.,

$$\tilde{f}(\rho, \theta) := f(x, y).$$

We shall also use the stretched variable $\hat{\rho} = \rho/t$. Circumflexes denote the corresponding change of variables

$$\hat{f}(\hat{\rho}, \theta) := \tilde{f}(\rho, \theta) = f(x, y).$$

3. An asymptotic expansion of the solution. We now turn to the construction of an asymptotic expansion with respect to the scaled thickness $t = \bar{t}/\sqrt{\lambda}$ of the solution of the Reissner–Mindlin clamped plate model using the formulation given in (2.3)–(2.7). Clearly r does not depend on t, so we begin, in this section, with the expansion of $\boldsymbol{\phi}$ and p. In § 6 we consider the expansion of ω and the shear stress.

Our immediate goal is to develop approximations of $\boldsymbol{\phi}$ and p by sums of the form

$$\boldsymbol{\phi}(x, y) \sim \boldsymbol{\phi}^I(x, y) + \boldsymbol{\phi}^B(x, y) := \sum_{i=0}^{\infty} t^i \boldsymbol{\phi}_i(x, y) + t^2 \tilde{\chi}(\rho) \sum_{i=0}^{\infty} t^i \hat{\boldsymbol{\Phi}}_i(\hat{\rho}, \theta),$$

$$p(x, y) \sim p^I(x, y) + p^B(x, y) := \sum_{i=0}^{\infty} t^i p_i(x, y) + t \tilde{\chi}(\rho) \sum_{i=0}^{\infty} t^i \hat{P}_i(\hat{\rho}, \theta),$$

where $\tilde{\chi}(\rho)$ is a smooth cutoff function which is identically one for $0 \leq \rho \leq \rho_0/3$ and which is identically zero for $\rho > 2\rho_0/3$. (The power of t multiplying the second sum in each expansion was chosen in anticipation of the results that follow.) In this section we shall calculate formally in order to motivate appropriate definitions of the interior expansion functions $\boldsymbol{\phi}_i$ and p_i, and the boundary correctors $\hat{\boldsymbol{\Phi}}_i$ and $\hat{P}_i$. In the next section we derive some estimates for these functions, and in § 5 we give rigorous bounds for the errors in the asymptotic expansions.

Now $\boldsymbol{\phi} \in \boldsymbol{H}^1(\Omega)$ and $p \in H^1(\Omega)/\mathbb{R}$ are uniquely determined by the equations

$$\begin{aligned} &-\mathbf{div}\, C\,\mathcal{E}(\boldsymbol{\phi}) - \mathbf{curl}\, p = \mathbf{grad}\, r \quad \text{in } \Omega, \\ &-\operatorname{rot} \boldsymbol{\phi} + t^2 \Delta p = 0 \ (\operatorname{mod} \mathbb{R}) \quad \text{in } \Omega, \\ &\boldsymbol{\phi} = 0, \quad \frac{\partial p}{\partial n} = 0 \quad \text{on } \partial\Omega. \end{aligned}$$

In writing the second equation modulo $\mathbb{R}$ we mean that $\boldsymbol{\phi}$ and p are to be determined with $-\operatorname{rot} \boldsymbol{\phi} + t^2 \Delta p$ equal to an unspecified constant function. In fact if we integrate this equation using the divergence theorem, it follows that if $\boldsymbol{\phi}$ and p also satisfy the boundary conditions, then the constant must vanish. Thus, although the equation modulo $\mathbb{R}$ is formally weaker than (2.5), in fact together with the other equation and the boundary conditions, we have an equivalent problem to (2.4), (2.5), (2.7). The reason for introducing this complication is that it is more convenient to define an asymptotic expansion that satisfies the second equation only up to an additive constant.

Formally, $(\boldsymbol{\phi}^I, p^I)$ will be determined such that

$$-\mathbf{div}\, C\,\mathcal{E}(\boldsymbol{\phi}^I) - \mathbf{curl}\, p^I = \mathbf{grad}\, r \quad \text{in } \Omega, \tag{3.1}$$

$$-\operatorname{rot} \boldsymbol{\phi}^I + t^2 \Delta p^I = 0 \ (\operatorname{mod} \mathbb{R}) \quad \text{in } \Omega, \tag{3.2}$$

$$\boldsymbol{\phi}^I = -\boldsymbol{\phi}^B \quad \text{on } \partial\Omega, \tag{3.3}$$

and $(\boldsymbol{\phi}^B, p^B)$ will be determined such that

$$-\mathbf{div}\, C\,\mathcal{E}(\boldsymbol{\phi}^B) - \mathbf{curl}\, p^B = 0 \quad \text{in } \Omega, \tag{3.4}$$

$$-\operatorname{rot} \boldsymbol{\phi}^B + t^2 \Delta p^B = 0 \quad \text{in } \Omega, \tag{3.5}$$

$$\frac{\partial p^B}{\partial n} = -\frac{\partial p^I}{\partial n} \quad \text{on } \partial\Omega. \tag{3.6}$$

Inserting the series expansions for $\boldsymbol{\phi}^I$, p^I, and $\boldsymbol{\phi}^B$ in the first boundary value problem and equating coefficients of corresponding powers of t, we obtain the boundary value problems defining the interior expansion functions $\boldsymbol{\phi}_i$ and p_i:

$$-\mathbf{div}\, C\,\mathcal{E}(\boldsymbol{\phi}_i) - \mathbf{curl}\, p_i = \begin{cases} \mathbf{grad}\, r & \text{for } i = 0, \\ 0 & \text{for } i > 0, \end{cases} \tag{3.7}$$

$$-\operatorname{rot} \boldsymbol{\phi}_i = \begin{cases} 0 \ (\operatorname{mod} \mathbb{R}) & \text{for } i = 0, 1, \\ -\Delta p_{i-2} \ (\operatorname{mod} \mathbb{R}) & \text{for } i \geq 2, \end{cases} \tag{3.8}$$

and the boundary conditions

$$\phi_i = \begin{cases} 0 & \text{for } i = 0, 1, \\ -\Phi_{i-2} & \text{for } i \geq 2. \end{cases} \tag{3.9}$$

In fact, (3.8) can be replaced by the simpler equation

$$\operatorname{rot} \phi_i = 0 \pmod{\mathbb{R}}. \tag{3.10}$$

To see this, apply rot to (3.7). Using simple calculus identities, we get that

$$-\frac{E}{24(1+\nu)} \Delta \operatorname{rot} \phi_i + \Delta p_i = 0.$$

It then follows from (3.8) that $\Delta p_i = 0$ for $i = 0, 1$ and, for $i \geq 2$,

$$\Delta p_i = \frac{E}{24(1+\nu)} \Delta^2 p_{i-2}.$$

By induction, $\Delta p_i = 0$ for all i. We thus use the system (3.7), (3.10), (3.9) to define the interior expansion functions. This system is essentially the Stokes equations and admits a unique solution in $\boldsymbol{H}^1(\Omega) \times L^2(\Omega)/\mathbb{R}$ (see Lemma 4.2 below). Note that the implied constant in (3.10) is uniquely determined by compatibility between this equation and the boundary conditions in (3.9). We also remark that the right-hand side of all three equations vanishes for $i = 1$, so $\phi_1 = 0$ and $p_1 = 0$.

To obtain the defining equations for the boundary correctors, we transform the system (3.4)–(3.6) to ρ-θ coordinates. The equation for $\tilde{\phi}^B$ and $\tilde{p}^B$ corresponding to (3.4) is

$$\tilde{\mathcal{A}}_0 \frac{\partial^2 \tilde{\phi}^B}{\partial \rho^2} + \tilde{\mathcal{A}}_1 \frac{\partial^2 \tilde{\phi}^B}{\partial \rho \partial \theta} + \tilde{\mathcal{A}}_2 \frac{\partial \tilde{\phi}^B}{\partial \rho} + \tilde{\mathcal{A}}_3 \frac{\partial^2 \tilde{\phi}^B}{\partial \theta^2} + \tilde{\mathcal{A}}_4 \frac{\partial \tilde{\phi}^B}{\partial \theta} + \tilde{\boldsymbol{A}}_5 \frac{\partial \tilde{p}^B}{\partial \rho} + \tilde{\boldsymbol{A}}_6 \frac{\partial \tilde{p}^B}{\partial \theta} = 0, \tag{3.11}$$

where

$$\mathcal{A}_0 = -D \begin{pmatrix} (\rho_x)^2 + (1-\nu)(\rho_y)^2/2 & (1+\nu)\rho_x\rho_y/2 \\ (1+\nu)\rho_x\rho_y/2 & (\rho_y)^2 + (1-\nu)(\rho_x)^2/2 \end{pmatrix},$$

$$\mathcal{A}_1 = -D \begin{pmatrix} 2\theta_x\rho_x + (1-\nu)\theta_y\rho_y & (1+\nu)(\theta_y\rho_x + \theta_x\rho_y)/2 \\ (1+\nu)(\theta_y\rho_x + \theta_x\rho_y)/2 & 2\theta_y\rho_y + (1-\nu)\theta_x\rho_x \end{pmatrix},$$

$$\mathcal{A}_2 = -D \begin{pmatrix} \rho_{xx} + (1-\nu)\rho_{yy}/2 & (1+\nu)\rho_{xy}/2 \\ (1+\nu)\rho_{xy}/2 & \rho_{yy} + (1-\nu)\rho_{xx}/2 \end{pmatrix},$$

$$\mathcal{A}_3 = -D \begin{pmatrix} (\theta_x)^2 + (1-\nu)(\theta_y)^2/2 & (1+\nu)\theta_x\theta_y/2 \\ (1+\nu)\theta_x\theta_y/2 & (\theta_y)^2 + (1-\nu)(\theta_x)^2/2 \end{pmatrix},$$

$$\mathcal{A}_4 = -D \begin{pmatrix} \theta_{xx} + (1-\nu)\theta_{yy}/2 & (1+\nu)\theta_{xy}/2 \\ (1+\nu)\theta_{xy}/2 & \theta_{yy} + (1-\nu)\theta_{xx}/2 \end{pmatrix},$$

$$\boldsymbol{A}_5 = -\operatorname{\mathbf{curl}} \rho, \quad \boldsymbol{A}_6 = -\operatorname{\mathbf{curl}} \theta.$$

Note that

$$\boldsymbol{A}_5 = \boldsymbol{s}, \quad \boldsymbol{A}_6 = \boldsymbol{n} \quad \text{on } \partial\Omega.$$

In ρ-θ coordinates (3.5) becomes

$$-\tilde{\boldsymbol{A}}_5 \cdot \frac{\partial \tilde{\boldsymbol{\phi}}^B}{\partial \rho} - \tilde{\boldsymbol{A}}_6 \cdot \frac{\partial \tilde{\boldsymbol{\phi}}^B}{\partial \theta} + t^2 \left(\frac{\partial^2 \tilde{p}^B}{\partial \rho^2} + \tilde{A}_7 \frac{\partial \tilde{p}^B}{\partial \rho} + \tilde{A}_8 \frac{\partial^2 \tilde{p}^B}{\partial \theta^2} + \tilde{A}_9 \frac{\partial \tilde{p}^B}{\partial \theta} \right) = 0, \tag{3.12}$$

where

$$A_7 = \Delta \rho, \quad A_8 = |\mathbf{grad}\, \theta|^2, \quad A_9 = \Delta \theta.$$

In deriving (3.12), we use the facts that $|\mathbf{grad}\, \rho| = 1$ and $\mathbf{grad}\, \rho \cdot \mathbf{grad}\, \theta = 0$. The boundary condition (3.6) becomes

$$\frac{\partial \tilde{p}^B}{\partial \rho} = \widetilde{\frac{\partial p^I}{\partial n}} \quad \text{on } \partial\Omega. \tag{3.13}$$

The exact form of the coefficient functions in (3.11) and (3.12) is not essential. However, the coefficients $\tilde{\mathcal{A}}_0$ and $\tilde{\boldsymbol{A}}_5$ have some properties which will prove important.

LEMMA 3.1. *The matrix-valued function $\tilde{\mathcal{A}}_0(\rho, \theta)$ and the vector-valued function $\tilde{\boldsymbol{A}}_5(\rho, \theta)$ are independent of ρ. Moreover for each fixed θ, $\tilde{\mathcal{A}}_0$ is symmetric negative definite and $\tilde{\boldsymbol{A}}_5$ is a unit eigenvector of $\tilde{\mathcal{A}}_0$ with eigenvalue $-D(1-\nu)/2$.*

Proof. That these coefficients are independent of ρ follows from the observation that $\widetilde{\partial\rho/\partial x}$ and $\widetilde{\partial\rho/\partial y}$ depend on θ but not on ρ. The second sentence is easily verified using the fact that $|\mathbf{grad}\, \rho| = 1$. □

The remaining coefficients are in general functions of both ρ and θ and to obtain the boundary layer equations, we expand them in Taylor series about $\rho = 0$. That is, we define operators $\mathcal{A}_i^j(\theta)$ by the formal Taylor series expansions:

$$\tilde{\mathcal{A}}_i(\rho, \theta) = \sum_{j=0}^{\infty} \frac{\rho^j}{j!} \tilde{\mathcal{A}}_i^j(\theta), \qquad i = 1, 2, 3, 4,$$

and define $\tilde{\boldsymbol{A}}_6^j$ and $\tilde{A}_i^j$, $i = 7, 8, 9$, similarly. Formally inserting these expansions in (3.11) and at the same time making the change of variable $\rho = t\hat{\rho}$ gives

$$\begin{aligned} &t^{-2} \hat{\mathcal{A}}_0 \frac{\partial^2 \hat{\boldsymbol{\phi}}^B}{\partial \hat{\rho}^2} + t^{-1} \left[\hat{\mathcal{A}}_1^0 \frac{\partial^2 \hat{\boldsymbol{\phi}}^B}{\partial \hat{\rho} \partial \theta} + \hat{\mathcal{A}}_2^0 \frac{\partial \hat{\boldsymbol{\phi}}^B}{\partial \hat{\rho}} + \hat{\boldsymbol{A}}_5 \frac{\partial \hat{p}^B}{\partial \hat{\rho}} \right] \\ &\quad + \sum_{j=0}^{\infty} t^j \left[\frac{\hat{\rho}^{j+1}}{(j+1)!} \left(\hat{\mathcal{A}}_1^{j+1} \frac{\partial^2 \hat{\boldsymbol{\phi}}^B}{\partial \hat{\rho} \partial \theta} + \hat{\mathcal{A}}_2^{j+1} \frac{\partial \hat{\boldsymbol{\phi}}^B}{\partial \hat{\rho}} \right) \right. \\ &\qquad \left. + \frac{\hat{\rho}^j}{j!} \left(\hat{\mathcal{A}}_3^j \frac{\partial^2 \hat{\boldsymbol{\phi}}^B}{\partial \theta^2} + \hat{\mathcal{A}}_4^j \frac{\partial \hat{\boldsymbol{\phi}}^B}{\partial \theta} + \hat{\boldsymbol{A}}_6^j \cdot \frac{\partial \hat{p}^B}{\partial \theta} \right) \right] = 0. \end{aligned} \tag{3.14}$$

Similarly (3.12) becomes:

$$\begin{aligned} &-t^{-1} \hat{\boldsymbol{A}}_5 \cdot \frac{\partial \hat{\boldsymbol{\phi}}^B}{\partial \hat{\rho}} - \hat{\boldsymbol{A}}_6^0 \cdot \frac{\partial \hat{\boldsymbol{\phi}}^B}{\partial \theta} + \frac{\partial^2 \hat{p}^B}{\partial \hat{\rho}^2} \\ &\quad + \sum_{j=0}^{\infty} t^{j+1} \left[-\frac{\hat{\rho}^{j+1}}{(j+1)!} \hat{\boldsymbol{A}}_6^{j+1} \cdot \frac{\partial \hat{\boldsymbol{\phi}}^B}{\partial \theta} + \frac{\hat{\rho}^j}{j!} \left(\hat{A}_7^j \frac{\partial \hat{p}^B}{\partial \hat{\rho}} + t \hat{A}_8^j \frac{\partial^2 \hat{p}^B}{\partial \theta^2} + t \hat{A}_9^j \frac{\partial \hat{p}^B}{\partial \theta} \right) \right] = 0. \end{aligned} \tag{3.15}$$

We now calculate the differential equations determining the boundary correctors by inserting the series expansions for ϕ^B and p^B (defined at the beginning of this section) in (3.14) and (3.15) and equating coefficients of corresponding powers of t. Neglecting the cutoff function χ, we obtain from (3.14) the equations

$$\hat{\mathcal{A}}_0 \frac{\partial^2 \hat{\boldsymbol{\Phi}}_0}{\partial \hat{\rho}^2} + \hat{\boldsymbol{A}}_5 \frac{\partial \hat{P}_0}{\partial \hat{\rho}} = 0,$$

$$\hat{\mathcal{A}}_0 \frac{\partial^2 \hat{\boldsymbol{\Phi}}_1}{\partial \hat{\rho}^2} + \hat{\boldsymbol{A}}_5 \frac{\partial \hat{P}_1}{\partial \hat{\rho}} + \hat{\mathcal{A}}_1^0 \frac{\partial^2 \hat{\boldsymbol{\Phi}}_0}{\partial \hat{\rho} \partial \theta} + \hat{\mathcal{A}}_2^0 \frac{\partial \hat{\boldsymbol{\Phi}}_0}{\partial \hat{\rho}} + \hat{\boldsymbol{A}}_6^0 \frac{\partial \hat{P}_0}{\partial \theta} = 0,$$

and, for $i = 2, 3, \cdots$,

$$\hat{\mathcal{A}}_0 \frac{\partial^2 \hat{\boldsymbol{\Phi}}_i}{\partial \hat{\rho}^2} + \hat{\boldsymbol{A}}_5 \frac{\partial \hat{P}_i}{\partial \hat{\rho}} + \hat{\mathcal{A}}_1^0 \frac{\partial^2 \hat{\boldsymbol{\Phi}}_{i-1}}{\partial \hat{\rho} \partial \theta} + \hat{\mathcal{A}}_2^0 \frac{\partial \hat{\boldsymbol{\Phi}}_{i-1}}{\partial \hat{\rho}} + \hat{\boldsymbol{A}}_6^0 \frac{\partial \hat{P}_{i-1}}{\partial \theta}$$
$$+ \sum_{j=0}^{i-2} \left[\frac{\hat{\rho}^{j+1}}{(j+1)!} \left(\hat{\mathcal{A}}_1^{j+1} \frac{\partial^2 \hat{\boldsymbol{\Phi}}_{i-2-j}}{\partial \hat{\rho} \partial \theta} + \hat{\mathcal{A}}_2^{j+1} \frac{\partial \hat{\boldsymbol{\Phi}}_{i-2-j}}{\partial \hat{\rho}} + \hat{\boldsymbol{A}}_6^{j+1} \frac{\partial \hat{P}_{i-2-j}}{\partial \theta} \right) \right.$$
$$\left. + \frac{\hat{\rho}^j}{j!} \left(\hat{\mathcal{A}}_3^j \frac{\partial^2 \hat{\boldsymbol{\Phi}}_{i-2-j}}{\partial \theta^2} + \hat{\mathcal{A}}_4^j \frac{\partial \hat{\boldsymbol{\Phi}}_{i-2-j}}{\partial \theta} \right) \right] = 0.$$

Introducing the convention $\hat{\boldsymbol{\Phi}}_i = 0$, $\hat{P}_i = 0$ for $i < 0$, we may write these three equations as

$$\hat{\mathcal{A}}_0 \frac{\partial^2 \hat{\boldsymbol{\Phi}}_i}{\partial \hat{\rho}^2} + \hat{\boldsymbol{A}}_5 \frac{\partial \hat{P}_i}{\partial \hat{\rho}} = -\hat{\boldsymbol{F}}_i(\hat{\rho}, \theta), \qquad i \in \mathbb{N}, \tag{3.16}$$

where

$$\hat{\boldsymbol{F}}_i(\hat{\rho}, \theta) = \sum_{j=0}^{i-1} \frac{\hat{\rho}^j}{j!} \left(\hat{\mathcal{A}}_1^j \frac{\partial^2 \hat{\boldsymbol{\Phi}}_{i-1-j}}{\partial \hat{\rho} \partial \theta} + \hat{\mathcal{A}}_2^j \frac{\partial \hat{\boldsymbol{\Phi}}_{i-1-j}}{\partial \hat{\rho}} + \hat{\mathcal{A}}_3^j \frac{\partial^2 \hat{\boldsymbol{\Phi}}_{i-2-j}}{\partial \theta^2} \right.$$
$$\left. + \hat{\mathcal{A}}_4^j \frac{\partial \hat{\boldsymbol{\Phi}}_{i-2-j}}{\partial \theta} + \hat{\boldsymbol{A}}_6^j \frac{\partial \hat{P}_{i-1-j}}{\partial \theta} \right).$$

Similarly, from (3.15), we obtain

$$-\hat{\boldsymbol{A}}_5 \cdot \frac{\partial \hat{\boldsymbol{\Phi}}_i}{\partial \hat{\rho}} + \frac{\partial^2 \hat{P}_i}{\partial \hat{\rho}^2} = \hat{G}_i(\hat{\rho}, \theta) :=$$
$$- \sum_{j=0}^{i-1} \frac{\hat{\rho}^j}{j!} \left(-\hat{\boldsymbol{A}}_6^j \cdot \frac{\partial \hat{\boldsymbol{\Phi}}_{i-1-j}}{\partial \theta} + \hat{A}_7^j \frac{\partial \hat{P}_{i-1-j}}{\partial \hat{\rho}} + \hat{A}_8^j \frac{\partial^2 \hat{P}_{i-2-j}}{\partial \theta^2} + \hat{A}_9^j \frac{\partial \hat{P}_{i-2-j}}{\partial \theta} \right), \tag{3.17}$$
$$i \in \mathbb{N}.$$

Inserting the asymptotic expansions for p^I and p^B in (3.13), changing variables from ρ to $\hat{\rho}$, and matching powers, we obtain the boundary conditions

$$\frac{\partial \hat{P}_i}{\partial \hat{\rho}}(0, \theta) = \widetilde{\frac{\partial p_i}{\partial n}}(0, \theta), \qquad i \in \mathbb{N}. \tag{3.18}$$

Finally, in order to determine the boundary correctors uniquely, we also impose the conditions at infinity

$$\lim_{\hat{\rho}\to\infty} \hat{\boldsymbol{\Phi}}_i(\hat{\rho},\theta) = 0, \quad \lim_{\hat{\rho}\to\infty} \hat{P}_i(\hat{\rho},\theta) = 0. \tag{3.19}$$

We remark that (3.17) is to be satisfied exactly, rather than up to an additive constant (as was (3.10)). Similarly, because of the boundary condition at infinity, we have specified $\hat{P}_i$ completely, not just up to an additive constant as was the case for p_i. Once the boundary correctors $\hat{\boldsymbol{\Phi}}_k$ and $\hat{P}_k$ for $k < i$ and the interior expansion function p_i are known, we may view (3.16)–(3.19) as a boundary value problem in ordinary differential equations in which the independent variable is $\hat{\rho}$, the unknowns are $\hat{\boldsymbol{\Phi}}_i$ and $\hat{P}_i$, and θ plays the role of a parameter. As we shall see (in Theorem 3.3), this problem has a unique solution. Therefore we can recursively determine all the interior expansion functions and boundary correctors as follows. First we determine $(\boldsymbol{\phi}_0, p_0)$ by (3.7), (3.10), and (3.9). Then we determine $(\hat{\boldsymbol{\Phi}}_0, \hat{P}_0)$ by (3.16)–(3.19) (the right-hand sides of (3.16) and (3.17) being zero and the right-hand side of (3.18) being known). Then $(\boldsymbol{\phi}_1, p_1)$ is uniquely determined by (3.7), (3.10), and (3.9), and so forth. Thus we have proved the following theorem.

THEOREM 3.2. *There exist functions $\boldsymbol{\phi}_i(x,y)$, $p_i(x,y)$ on Ω and $\hat{\boldsymbol{\Phi}}_i(\hat{\rho},\theta)$, $\hat{P}_i(\hat{\rho},\theta)$ on $\hat{\Omega}_0$, $i \in \mathbb{N}$, unique except that p_i is determined only up to an additive constant, which satisfy the boundary value problems* (3.7), (3.10), (3.9) *and* (3.16)–(3.19).

The Stokes-like boundary value problem (3.7), (3.10), (3.9) is well posed, but, of course, we cannot in general determine its solution in closed form, even if r were known in closed form. (However, the regularity of solutions to this problem is well understood—cf. Lemma 4.2.) The system (3.16)–(3.19) can, in principle, be solved in closed form. For example, the solution for $i = 0$ is

$$\hat{P}_0(\hat{\rho},\theta) = -\frac{1}{c}\widehat{\frac{\partial p_0}{\partial n}}(0,\theta)e^{-c\hat{\rho}}, \qquad \hat{\boldsymbol{\Phi}}_0(\hat{\rho},\theta) = \hat{\boldsymbol{A}}_5(\theta)\widehat{\frac{\partial p_0}{\partial n}}(0,\theta)e^{-c\hat{\rho}}, \tag{3.20}$$

where $c = [24(1+\nu)/E]^{1/2}$. (We show that this is the only solution in the proof of Theorem 3.3.)

The following theorem gives the form of the solution for general i. In particular, it states that $\hat{\boldsymbol{\Phi}}_i$ and $\hat{P}_i$ are polynomials in $\hat{\rho}$ times the decaying exponential $e^{-c\hat{\rho}}$.

THEOREM 3.3. *For each $i \in \mathbb{N}$, the system* (3.16)–(3.19) *has a unique solution $(\hat{\boldsymbol{\Phi}}_i, \hat{P}_i)$. Moreover there exist smooth functions $\boldsymbol{\alpha}_{ijkl}(\theta)$ and $\alpha_{ijkl}(\theta)$ depending only on i, the domain Ω, and the plate constants E and ν such that*

$$\hat{\boldsymbol{\Phi}}_i(\hat{\rho},\theta) = e^{-c\hat{\rho}} \sum_{k=0}^{i}\sum_{j=0}^{i}\sum_{l=0}^{i-j} \boldsymbol{\alpha}_{ijkl}(\theta)\hat{\rho}^k \frac{\partial^l}{\partial\theta^l}\widehat{\frac{\partial p_j}{\partial n}}(0,\theta),$$

$$\hat{P}_i(\hat{\rho},\theta) = e^{-c\hat{\rho}} \sum_{k=0}^{i}\sum_{j=0}^{i}\sum_{l=0}^{i-j} \alpha_{ijkl}(\theta)\hat{\rho}^k \frac{\partial^l}{\partial\theta^l}\widehat{\frac{\partial p_j}{\partial n}}(0,\theta).$$

The proof, an exercise in ordinary differential equations based on the form of the coefficients of (3.16)–(3.17) as given in Lemma 3.1, is given in the Appendix.

This completes the construction of the interior and boundary layer asymptotic expansions. In the next section we bound the individual terms of the series and determine how nearly the finite sums satisfy the differential systems which motivated their definitions. Then, in § 5, we prove error bounds for the finite sums of the expansions.

4. A priori estimates. We begin this section by deriving a priori bounds on the boundary correctors using Theorem 3.3.

THEOREM 4.1 (A PRIORI ESTIMATES FOR BOUNDARY CORRECTORS). *Let i be a nonnegative integer. There exists a constant C depending only on the domain Ω, the elastic constants E and ν, and s and i, such that*

$$|\boldsymbol{\Phi}_i|_s + |P_i|_s \le C \sum_{j=0}^{i} \left| \frac{\partial p_j}{\partial n} \right|_{s+i-j}, \qquad s \in \mathbb{R},$$

$$\|\boldsymbol{\Phi}_i\|_{s,\Omega_0} + \|P_i\|_{s,\Omega_0} \le C t^{1/2-s} \sum_{j=0}^{i} \sum_{m=0}^{s} t^m \left| \frac{\partial p_j}{\partial n} \right|_{m+i-j}, \qquad s \in \mathbb{N}.$$

Proof. The first estimate follows from Theorem 3.3 by setting $\hat{\rho} = 0$ and using the triangle inequality. We now consider the second inequality. To establish the bound for $\boldsymbol{\Phi}_i$, we change to (ρ, θ) coordinates and seek bounds on the integrals

(4.1)
$$\left[\int_0^L \int_0^{\rho_0} |\partial^{s-m+k} \tilde{\boldsymbol{\Phi}}_i / \partial \rho^{s-m} \partial \theta^k|^2 |1 - \kappa(\theta)\rho| \, d\rho \, d\theta \right]^{1/2}, \qquad 0 \le m \le s, \quad 0 \le k \le m.$$

Now $\tilde{\boldsymbol{\Phi}}_i$ is a sum of terms of the form

(4.2)
$$\boldsymbol{\alpha}(\theta) \exp(-c\rho/t) f(\rho/t) \frac{\partial^l}{\partial \theta^l} \widetilde{\frac{\partial p_j}{\partial n}}(0, \theta), \qquad \boldsymbol{\alpha} \text{ smooth, } f \text{ polynomial, } j \le i, \; l \le i - j.$$

The $L^2(\Omega_0)$ norm of (4.2) is bounded by $Ct^{1/2} |\partial p_j / \partial n|_l$, since

$$\int_0^{\rho_0} |\exp(-c\rho/t) f(\rho/t)|^2 \, d\rho \le t \int_0^{\infty} |\exp(-c\hat{\rho}) f(\hat{\rho})|^2 \, d\hat{\rho}.$$

Applying $\partial^{s-m+k} / \partial \rho^{s-m} \partial \theta^k$ to (4.2) gives t^{m-s} times a sum of terms of the same form except that l may be as large as $i - j + k \le m + i - j$. Thus (4.1) is bounded by $Ct^{1/2} \sum_{j=0}^{i} t^{m-s} |\partial p_j / \partial n|_{m+i-j}$. Summing over $m = 0, 1, \cdots, s$ gives the desired bound for $\tilde{\boldsymbol{\Phi}}_i$, and that for P_i is proved identically. □

We next summarize the basic regularity properties of the Stokes-like system which defines the interior expansion functions.

LEMMA 4.2. *Let $s \in \mathbb{N}$, $f \in H^s(\Omega) \cap \mathring{H}^1(\Omega)$, $g \in H^s(\Omega)/\mathbb{R}$, and $\boldsymbol{l} \in \boldsymbol{H}^{s+1/2}(\partial\Omega)$ be given. Then there exist unique $\boldsymbol{\psi} \in \boldsymbol{H}^{s+1}(\Omega)$, $q \in H^s(\Omega)/\mathbb{R}$ satisfying the partial differential equations*

$$-\mathbf{div}\, C\, \mathcal{E}(\boldsymbol{\psi}) - \mathbf{curl}\, q = \mathbf{grad}\, f, \tag{4.3}$$

$$-\operatorname{rot} \boldsymbol{\psi} = g \pmod{\mathbb{R}}, \tag{4.4}$$

and the boundary conditions

$$\boldsymbol{\psi} = \boldsymbol{l}.$$

Moreover, there exists a constant C depending only on s, E, ν, and Ω such that

$$\|\boldsymbol{\psi}\|_{s+1} + \|q\|_{s/\mathbb{R}} \leq C(\|f\|_s + \|g\|_{s/\mathbb{R}} + |\boldsymbol{l}|_{s+1/2}).$$

Remarks. 1. The restriction that the forcing function in (4.3) be the gradient of an $\mathring{H}^1(\Omega)$ function is sufficient for our purposes and allows us to avoid some technical points concerning duals of Sobolev spaces and trace operators with values in negative order spaces. 2. If we replace $\mathbf{div}\, C\,\mathcal{E}(\boldsymbol{\psi})$ with Δ then the simple change of variables $(\psi_1, \psi_2) \to (\psi_2, -\psi_1)$ converts (4.3), (4.4) to a generalized Stokes system, and this result is well known [5]. Here we give a proof which works for general C based on regularity results for the biharmonic.

Proof. Written in weak form, the boundary value problem is to find $\boldsymbol{\psi} \in \boldsymbol{H}^1(\Omega)$ such that $\boldsymbol{\psi} = \boldsymbol{l}$ on $\partial\Omega$ and $q \in L^2(\Omega)/\mathbb{R}$ satisifying

$$\begin{aligned}(C\,\mathcal{E}(\boldsymbol{\psi}), \mathcal{E}(\boldsymbol{\mu})) - (q, \operatorname{rot} \boldsymbol{\mu}) &= -(f, \operatorname{div} \boldsymbol{\mu}) \\ -(\operatorname{rot} \boldsymbol{\psi}, v) &= (g, v)\end{aligned}$$

for all $\boldsymbol{\mu} \in \mathring{H}^1(\Omega)$ and all $v \in L^2(\Omega)$ of mean value zero. Existence and uniqueness of a solution in $\boldsymbol{H}^1(\Omega) \times L^2(\Omega)/\mathbb{R}$ is proved just as for the generalized Stokes equations, e.g., by applying Brezzi's theorem [4]. The estimate for $s = 0$ follows from the same argument. To establish the estimate with $s \geq 1$, we apply the Helmholtz decomposition to $\boldsymbol{\psi}$ to get $\boldsymbol{\psi} = \mathbf{grad}\, z + \mathbf{curl}\, b$, with $z \in \mathring{H}^1(\Omega)$, $b \in H^1(\Omega)/\mathbb{R}$. From (4.4) we get

$$\Delta b = g \pmod{\mathbb{R}} \quad \text{in } \Omega,$$

with boundary conditions $\partial b/\partial n = \boldsymbol{l} \cdot \boldsymbol{s}$. Taking the divergence of equation (4.3) gives

$$-D\,\Delta^2 z = \Delta f \quad \text{in } \Omega,$$

with boundary conditions $z = 0$, $\partial z/\partial n = \boldsymbol{l} \cdot \boldsymbol{n} + \partial b/\partial s$. Applying regularity results for the biharmonic problem and the Laplacian, we obtain

$$\begin{aligned}\|b\|_{s+2/\mathbb{R}} &\leq C(\|g\|_{s/\mathbb{R}} + + |\boldsymbol{l}|_{s+1/2}), \\ \|z\|_{s+2} &\leq C(\|\,\Delta f\|_{s-2} + |\boldsymbol{l}|_{s+1/2} + |\partial b/\partial s|_{s+1/2}) \\ &\leq C(\|f\|_s + |\boldsymbol{l}|_{s+1/2} + \|b\|_{s+2/\mathbb{R}}) \\ &\leq C(\|f\|_s + |\boldsymbol{l}|_{s+1/2} + \|g\|_{s/\mathbb{R}}).\end{aligned}$$

The bound for $\boldsymbol{\psi}$ now follows directly by the triangle inequality and the bound for q then follows from (4.3). □

We now use the previous two theorems to obtain estimates for the interior expansion functions.

THEOREM 4.3 (A PRIORI ESTIMATES FOR INTERIOR EXPANSION FUNCTIONS). *Let $\boldsymbol{\phi}_i$ and p_i be the interior expansion functions. Then for all $s \geq 0$ and $i \in \mathbb{N}$, there exists a constant C such that*

$$\|\boldsymbol{\phi}_i\|_{s+1} + \|p_i\|_{s/\mathbb{R}} \leq C|||g|||_{s+i-2}.$$

Proof. Since $-\Delta r = g$ and r vanishes on $\partial\Omega$, we have

$$\|r\|_s \leq C|||g|||_{s-2}. \tag{4.5}$$

Thus, it suffices to prove that

$$\|\phi_i\|_{s+1} + \|p_i\|_{s/\mathbb{R}} \le C\|r\|_{s+i}. \tag{4.6}$$

We prove this first for $s \in \mathbb{N}$ by induction on i. For $i = 0$, we get this immediately from the defining equations (3.7), (3.10), (3.9), and Lemma 4.2. As already noted $\phi_1 = p_1 = 0$, so (4.6) holds for $i = 1$ also. For $i \ge 2$, we apply Lemma 4.2, Theorem 4.1, and the trace theorem to obtain

$$\|\phi_i\|_{s+1} + \|p_i\|_{s/\mathbb{R}} \le C|\boldsymbol{\Phi}_{i-2}|_{s+1/2} \le C\sum_{j=0}^{i-2}\left|\frac{\partial p_j}{\partial n}\right|_{s-3/2+i-j} \le C\sum_{j=0}^{i-2}\|p_j\|_{s+i-j}.$$

Application of the inductive hypothesis completes the proof of (4.6) for integer s. The proof for noninteger s now follows by a standard interpolation argument. □

COROLLARY 4.4. *For $s \ge -\frac{3}{2}$ and $i \in \mathbb{N}$, there exists a constant C such that*

$$\left|\frac{\partial p_i}{\partial s}\right|_s + \left|\frac{\partial p_i}{\partial n}\right|_s + |\boldsymbol{\Phi}_i|_s + |P_i|_s \le C|||g|||_{s+i-1/2}.$$

Proof. As remarked in the previous section, p_i is harmonic for all i. Consequently, the trace inequality

$$\left|\frac{\partial p_i}{\partial s}\right|_s + \left|\frac{\partial p_i}{\partial n}\right|_s \le C\|p_i\|_{s+3/2/\mathbb{R}}$$

holds for all s, so the bounds on p_i follow easily from the theorem. The bounds on $\boldsymbol{\Phi}_i$ and P_i then follow from Theorem 4.1. □

We now combine Theorems 4.1 and 4.3 to obtain an essential result for the derivation of error bounds for the boundary layer expansion.

THEOREM 4.5 (A PRIORI ESTIMATES FOR BOUNDARY CORRECTORS). *Let i, k, l, n, and s be nonnegative integers and define functions $\boldsymbol{f}$ and f on Ω_0 by*

$$\tilde{\boldsymbol{f}} = \left(\frac{\rho}{t}\right)^k t^l \frac{\partial^{l+n}}{\partial\rho^l\partial\theta^n}\tilde{\boldsymbol{\Phi}}_i, \qquad \tilde{f} = \left(\frac{\rho}{t}\right)^k t^l \frac{\partial^{l+n}}{\partial\rho^l\partial\theta^n}\tilde{P}_i.$$

Then there exists a constant C depending only on the domain Ω, the elastic constants E and ν, and i, k, l, n, and s, such that

$$\|\boldsymbol{f}\|_{s,\Omega_0} + \|f\|_{s,\Omega_0} \le C(t^{1/2-s}|||g|||_{i+n-1/2} + t^{1/2}|||g|||_{s+i+n-1/2}).$$

Proof. Note that, if in (4.2) we differentiate with respect to $\hat{\rho}$ (i.e., differentiate with respect to ρ and multiply by t), we obtain something of the same form. The same is true if we multiply by $\hat{\rho}$. If we differentiate with respect to θ we obtain a sum of two terms of the same form, with one higher order of differentiation on $\widetilde{\partial p_j/\partial n}$. Hence, reasoning just as in the proof of Theorem 4.1, we obtain

$$\|f\|_{s,\Omega_0} + \|\boldsymbol{f}\|_{s,\Omega_0} \le Ct^{1/2-s}\sum_{j=0}^{i}\sum_{m=0}^{s} t^m\left|\frac{\partial p_j}{\partial n}\right|_{m+i+n-j}.$$

An application of Corollary 4.4 completes the proof. □

We now consider the partial sums given by

$$\phi_m^I(x,y) = \sum_{i=0}^{m} t^i \phi_i(x,y), \qquad p_m^I(x,y) = \sum_{i=0}^{m} t^i p_i(x,y),$$

$$\phi_m^B(x,y) = t^2 \tilde{\chi}(\rho) \sum_{i=0}^{m} t^i \hat{\Phi}_i(\hat{\rho},\theta), \quad p_m^B(x,y) = t\tilde{\chi}(\rho) \sum_{i=0}^{m} t^i \hat{P}_i(\hat{\rho},\theta).$$

Note that while Φ_i and P_i are only defined on the tubular neighborhood Ω_0 of $\partial\Omega$, ϕ_m^B and p_m^B are defined on all of Ω because of the cutoff function χ. By construction, (ϕ_m^I, p_m^I) and (ϕ_m^B, p_m^B) should almost satisfy the boundary value problems (3.1)–(3.3) and (3.4)–(3.6), respectively. We now make precise to what extent this is true.

For the interior expansion this is easy. The following theorem follows directly from (3.7), (3.10), (3.9).

THEOREM 4.6 (BOUNDARY VALUE PROBLEM FOR THE INTERIOR EXPANSION). *Let $m \in \mathbb{N}$. The finite interior expansion $(\phi_m^I, p_m^I) \in \boldsymbol{H}^1(\Omega) \times H^1(\Omega)/\mathbb{R}$ satisfies the boundary value problem*

$$\begin{aligned} &-\mathbf{div}\, C\,\mathcal{E}(\phi_m^I) - \mathbf{curl}\, p_m^I = \mathbf{grad}\, r \quad \text{in } \Omega, \\ &-\operatorname{rot} \phi_m^I + t^2 \Delta p_m^I = 0 \ (\operatorname{mod} \mathbb{R}) \quad \text{in } \Omega, \\ &\phi_m^I = -\phi_{m-2}^B \quad \text{on } \partial\Omega. \end{aligned}$$

For the boundary expansion, it follows from (3.18) that the boundary condition

$$\frac{\partial p_m^B}{\partial n} = -\frac{\partial p_m^I}{\partial n} \quad \text{on } \partial\Omega \tag{4.7}$$

is satisfied exactly, but for the differential equations the situation is more complicated. Define the residuals $\boldsymbol{R}_m$ and R_m by the equations:

$$-\mathbf{div}\, C\,\mathcal{E}(\phi_m^B) - \mathbf{curl}\, p_m^B = \boldsymbol{R}_m \quad \text{in } \Omega, \tag{4.8}$$

$$-\operatorname{rot} \phi_m^B + t^2 \Delta p_m^B = R_m \quad \text{in } \Omega. \tag{4.9}$$

The following theorem shows that these residuals are indeed of high order with respect to t.

THEOREM 4.7 (BOUNDARY VALUE PROBLEM FOR THE BOUNDARY LAYER EXPANSION). *Let $m \in \mathbb{N}$. The finite boundary layer expansion $(\phi_m^B, p_m^B) \in \boldsymbol{H}^1(\Omega) \times H^1(\Omega)$ satisfies the boundary value problem (4.7)–(4.9), with the following bounds valid for the forcing functions $\boldsymbol{R}_m$, R_m:*

$$\begin{aligned} &\|\boldsymbol{R}_m\|_s \le C(t^{m+3/2-s}\|g\|_{m+1/2} + t^{m+5/2}\|g\|_{m+s+3/2}), \qquad s = -1, 0, \cdots, \\ &\|R_m\|_s \le C(t^{m+5/2-s}\|g\|_{m+1/2} + t^{m+7/2}\|g\|_{m+s+3/2}), \qquad s \in \mathbb{N}. \end{aligned}$$

Proof. It suffices to prove the theorem with the right-hand sides replaced by

$$Ct^{m+3/2-s} \sum_{i=0}^{s+1} t^i \|g\|_{m+1/2+i} \text{ and } Ct^{m+5/2-s} \sum_{i=0}^{s+1} t^i \|g\|_{m+1/2+i},$$

respectively. In addition to the portion of the residual due to truncating the series after finitely many terms, we must consider the contributions from two other sources, namely the replacement of the coefficients by Taylor polynomial approximations and the suppression of the cutoff function χ. Because of the presence of the cutoff function χ in the definitions of ϕ_m^B and p_m^B, it is easy to see that the residuals $\boldsymbol{R}_m$ and R_m will vanish for $\rho \geq \rho_0$, i.e., in $\Omega \setminus \Omega_0$. In Ω_0, after changing to $(\hat\rho, \theta)$ variables, we have (cf. (3.11))

$$\hat{\boldsymbol{R}}_m = t^{-2}\hat{\boldsymbol{A}}_0 \frac{\partial^2 \hat\phi_m^B}{\partial \hat\rho^2} + t^{-1}\left(\hat{\boldsymbol{A}}_1 \frac{\partial^2 \hat\phi_m^B}{\partial \hat\rho \partial\theta} + \hat{\boldsymbol{A}}_2 \frac{\partial \hat\phi_m^B}{\partial \hat\rho}\right) + \hat{\boldsymbol{A}}_3 \frac{\partial^2 \hat\phi_m^B}{\partial\theta^2} + \hat{\boldsymbol{A}}_4 \frac{\partial \hat\phi_m^B}{\partial\theta}$$
$$+ t^{-1}\hat{\boldsymbol{A}}_5 \frac{\partial \hat p_m^B}{\partial \hat\rho} + \hat{\boldsymbol{A}}_6 \frac{\partial \hat p_m^B}{\partial\theta} = \tilde\chi(\rho)\hat{\boldsymbol{R}}_m^1 + \hat{\boldsymbol{R}}_m^2,$$

where

$$\hat{\boldsymbol{R}}_m^1 = \sum_{i=0}^{m} t^i \left(\hat{\boldsymbol{A}}_0 \frac{\partial^2 \hat{\boldsymbol{\Phi}}_i}{\partial \hat\rho^2} + \hat{\boldsymbol{A}}_5 \frac{\partial \hat P_i}{\partial \hat\rho}\right)$$
$$+ \sum_{i=0}^{m+1} t^i \left(\hat{\boldsymbol{A}}_1 \frac{\partial^2 \hat{\boldsymbol{\Phi}}_{i-1}}{\partial\hat\rho\partial\theta} + \hat{\boldsymbol{A}}_2 \frac{\partial \hat{\boldsymbol{\Phi}}_{i-1}}{\partial\hat\rho} + \hat{\boldsymbol{A}}_3 \frac{\partial^2 \hat{\boldsymbol{\Phi}}_{i-2}}{\partial\theta^2} + \hat{\boldsymbol{A}}_4 \frac{\partial \hat{\boldsymbol{\Phi}}_{i-2}}{\partial\theta} + \hat{\boldsymbol{A}}_6 \frac{\partial \hat P_{i-1}}{\partial\theta}\right)$$
$$+ t^{m+2}\left(\hat{\boldsymbol{A}}_3 \frac{\partial^2 \hat{\boldsymbol{\Phi}}_m}{\partial\theta^2} + \hat{\boldsymbol{A}}_4 \frac{\partial \hat{\boldsymbol{\Phi}}_m}{\partial\theta}\right),$$
$$\hat{\boldsymbol{R}}_m^2 = \tilde\chi'(\rho)\left[\sum_{i=0}^{m} t^i \left(2\hat{\boldsymbol{A}}_0 \frac{\partial \hat{\boldsymbol{\Phi}}_i}{\partial\hat\rho} + \hat{\boldsymbol{A}}_5 \hat P_i\right) + \sum_{i=0}^{m+1} t^i \left(\hat{\boldsymbol{A}}_1 \frac{\partial \hat{\boldsymbol{\Phi}}_{i-1}}{\partial\theta} + \hat{\boldsymbol{A}}_2 \hat{\boldsymbol{\Phi}}_{i-1}\right)\right]$$
$$+ \tilde\chi''(\rho) \sum_{i=0}^{m} t^i \hat{\boldsymbol{A}}_0 \hat{\boldsymbol{\Phi}}_i,$$

and we have again used the convention that terms with negative indices vanish. Now for any $k \geq -1$

$$\tilde{\boldsymbol{A}}_1(\rho,\theta) = \sum_{j=0}^{k} \frac{\rho^j}{j!} \tilde{\boldsymbol{A}}_1^j(\theta) + \rho^{k+1} \bar{\boldsymbol{A}}_1^{k+1}(\rho,\theta)$$

where

$$\bar{\boldsymbol{A}}_1^{k+1}(\rho,\theta) = \begin{cases} \displaystyle\int_0^1 \frac{\partial^{k+1}\tilde{\boldsymbol{A}}_1}{\partial\rho^{k+1}}(s\rho,\theta)\frac{(1-s)^k}{k!}\,ds, & k \geq 0, \\ \tilde{\boldsymbol{A}}_1(\rho,\theta), & k = -1. \end{cases}$$

The other coefficients admit similar Taylor expansions (except for $\tilde{\boldsymbol{A}}_0$ and $\tilde{\boldsymbol{A}}_5$ which are functions of θ only). Substituting these expansions for $k = m - i$ and using $\rho = t\hat\rho$, we get

$$\hat{\boldsymbol{R}}_m^1 = \sum_{i=0}^{m} t^i \left(\hat{\boldsymbol{A}}_0 \frac{\partial^2 \hat{\boldsymbol{\Phi}}_i}{\partial\hat\rho^2} + \hat{\boldsymbol{A}}_5 \frac{\partial \hat P_i}{\partial\hat\rho}\right) + \sum_{i=0}^{m+1} t^i \sum_{j=0}^{m-i} \frac{(t\hat\rho)^j}{j!}\left(\hat{\boldsymbol{A}}_1^j \frac{\partial^2 \hat{\boldsymbol{\Phi}}_{i-1}}{\partial\hat\rho\partial\theta} + \hat{\boldsymbol{A}}_2^j \frac{\partial \hat{\boldsymbol{\Phi}}_{i-1}}{\partial\hat\rho}\right.$$
$$\left. + \hat{\boldsymbol{A}}_3^j \frac{\partial^2 \hat{\boldsymbol{\Phi}}_{i-2}}{\partial\theta^2} + \hat{\boldsymbol{A}}_4^j \frac{\partial \hat{\boldsymbol{\Phi}}_{i-2}}{\partial\theta} + \hat{\boldsymbol{A}}_6^j \frac{\partial \hat P_{i-1}}{\partial\theta}\right)$$

$$
\begin{aligned}
&+\sum_{i=0}^{m+1} t^i(t\hat{\rho})^{m-i+1}\left(\hat{\mathcal{A}}_1^{m-i+1}\frac{\partial^2\hat{\boldsymbol{\Phi}}_{i-1}}{\partial\hat{\rho}\partial\theta}+\hat{\mathcal{A}}_2^{m-i+1}\frac{\partial\hat{\boldsymbol{\Phi}}_{i-1}}{\partial\hat{\rho}}\right.\\
&\qquad\left.+\hat{\mathcal{A}}_3^{m-i+1}\frac{\partial^2\hat{\boldsymbol{\Phi}}_{i-2}}{\partial\theta^2}+\hat{\mathcal{A}}_4^{m-i+1}\frac{\partial\hat{\boldsymbol{\Phi}}_{i-2}}{\partial\theta}+\hat{\boldsymbol{A}}_6^{m-i+1}\frac{\partial\hat{P}_{i-1}}{\partial\theta}\right)\\
&+t^{m+2}\left(\hat{\mathcal{A}}_3\frac{\partial^2\hat{\boldsymbol{\Phi}}_m}{\partial\theta^2}+\hat{\mathcal{A}}_4\frac{\partial\hat{\boldsymbol{\Phi}}_m}{\partial\theta}\right)\\
=&\sum_{i=0}^{m} t^i\left[\hat{\mathcal{A}}_0\frac{\partial^2\hat{\boldsymbol{\Phi}}_i}{\partial\hat{\rho}^2}+\hat{\boldsymbol{A}}_5\frac{\partial\hat{P}_i}{\partial\hat{\rho}}+\sum_{j=0}^{i}\frac{\hat{\rho}^j}{j!}\left(\hat{\mathcal{A}}_1^j\frac{\partial^2\hat{\boldsymbol{\Phi}}_{i-j-1}}{\partial\hat{\rho}\partial\theta}+\hat{\mathcal{A}}_2^j\frac{\partial\hat{\boldsymbol{\Phi}}_{i-j-1}}{\partial\hat{\rho}}\right.\right.\\
&\qquad\left.\left.+\hat{\mathcal{A}}_3^j\frac{\partial^2\hat{\boldsymbol{\Phi}}_{i-j-2}}{\partial\theta^2}+\hat{\mathcal{A}}_4^j\frac{\partial\hat{\boldsymbol{\Phi}}_{i-j-2}}{\partial\theta}+\hat{\boldsymbol{A}}_6^j\frac{\partial\hat{P}_{i-j-1}}{\partial\theta}\right)\right]\\
&+t^{m+1}\sum_{i=0}^{m+1}\hat{\rho}^{m-i+1}\left(\hat{\mathcal{A}}_1^{m-i+1}\frac{\partial^2\hat{\boldsymbol{\Phi}}_{i-1}}{\partial\hat{\rho}\partial\theta}+\hat{\mathcal{A}}_2^{m-i+1}\frac{\partial\hat{\boldsymbol{\Phi}}_{i-1}}{\partial\hat{\rho}}\right.\\
&\qquad\left.+\hat{\mathcal{A}}_3^{m-i+1}\frac{\partial^2\hat{\boldsymbol{\Phi}}_{i-2}}{\partial\theta^2}+\hat{\mathcal{A}}_4^{m-i+1}\frac{\partial\hat{\boldsymbol{\Phi}}_{i-2}}{\partial\theta}+\hat{\boldsymbol{A}}_6^{m-i+1}\frac{\partial\hat{P}_{i-1}}{\partial\theta}\right)\\
&+t^{m+2}\left(\hat{\mathcal{A}}_3\frac{\partial^2\hat{\boldsymbol{\Phi}}_m}{\partial\theta^2}+\hat{\mathcal{A}}_4\frac{\partial\hat{\boldsymbol{\Phi}}_m}{\partial\theta}\right),
\end{aligned}
$$

where we have used the identity $\sum_{i=0}^{m}\sum_{j=0}^{m-i}F(i,j)=\sum_{i=0}^{m}\sum_{j=0}^{i}F(i-j,j)$ to obtain the second equality. Now the term in brackets vanishes by construction (cf. (3.16)). Thus

$$
\begin{aligned}
\hat{\boldsymbol{R}}_m^1=&\,t^{m+1}\sum_{i=0}^{m+1}\hat{\rho}^{m-i+1}\left(\hat{\mathcal{A}}_1^{m-i+1}\frac{\partial^2\hat{\boldsymbol{\Phi}}_{i-1}}{\partial\hat{\rho}\partial\theta}+\hat{\mathcal{A}}_2^{m-i+1}\frac{\partial\hat{\boldsymbol{\Phi}}_{i-1}}{\partial\hat{\rho}}\right.\\
&\left.+\hat{\mathcal{A}}_3^{m-i+1}\frac{\partial^2\hat{\boldsymbol{\Phi}}_{i-2}}{\partial\theta^2}+\hat{\mathcal{A}}_4^{m-i+1}\frac{\partial\hat{\boldsymbol{\Phi}}_{i-2}}{\partial\theta}+\hat{\boldsymbol{A}}_6^{m-i+1}\frac{\partial\hat{P}_{i-1}}{\partial\theta}\right)\\
&+t^{m+2}\left(\hat{\mathcal{A}}_3\frac{\partial^2\hat{\boldsymbol{\Phi}}_m}{\partial\theta^2}+\hat{\mathcal{A}}_4\frac{\partial\hat{\boldsymbol{\Phi}}_m}{\partial\theta}\right),
\end{aligned}
$$

or

$$
\begin{aligned}
\hat{\boldsymbol{R}}_m^1=&\,t^{m+1}\left[\sum_{j=0}^{m}\hat{\rho}^j\left(\hat{\mathcal{A}}_1^j\frac{\partial^2\hat{\boldsymbol{\Phi}}_{m-j}}{\partial\hat{\rho}\partial\theta}+\hat{\mathcal{A}}_2^j\frac{\partial\hat{\boldsymbol{\Phi}}_{m-j}}{\partial\hat{\rho}}+\hat{\mathcal{A}}_3^j\frac{\partial^2\hat{\boldsymbol{\Phi}}_{m-j-1}}{\partial\theta^2}\right.\right.\\
&\left.\left.+\hat{\mathcal{A}}_4^j\frac{\partial\hat{\boldsymbol{\Phi}}_{m-j-1}}{\partial\theta}+\hat{\boldsymbol{A}}_6^j\frac{\partial\hat{P}_{m-j-1}}{\partial\theta}\right)+t\left(\hat{\mathcal{A}}_3\frac{\partial^2\hat{\boldsymbol{\Phi}}_m}{\partial\theta^2}+\hat{\mathcal{A}}_4\frac{\partial\hat{\boldsymbol{\Phi}}_m}{\partial\theta}\right)\right].
\end{aligned}
\tag{4.10}
$$

A similar computation gives $\hat{R}_m=\tilde{\chi}(\rho)\hat{R}_m^1+\hat{R}_m^2$, where

$$
\begin{aligned}
\hat{R}_m^1=&\,t^{m+2}\left[\sum_{j=0}^{m}\hat{\rho}^j\left(-\hat{\boldsymbol{A}}_6^j\cdot\frac{\partial\hat{\boldsymbol{\Phi}}_{m-j}}{\partial\theta}+\hat{\mathcal{A}}_7^j\frac{\partial\hat{P}_{m-j}}{\partial\hat{\rho}}+\hat{\mathcal{A}}_8^j\frac{\partial^2\hat{P}_{m-j-1}}{\partial\theta^2}\right.\right.\\
&\left.\left.+\hat{\mathcal{A}}_9^j\frac{\partial\hat{P}_{m-j-1}}{\partial\theta}\right)+t\left(\hat{\mathcal{A}}_8\frac{\partial^2\hat{P}_m}{\partial\theta^2}+\hat{\mathcal{A}}_9\frac{\partial\hat{P}_m}{\partial\theta}\right)\right]
\end{aligned}
$$

and

$$\hat{R}_m^2 = \tilde{\chi}'(\rho)\left[t\sum_{i=0}^{m} t^i\left(-\hat{A}_5\hat{\boldsymbol{\Phi}}_i + 2\frac{\partial \hat{P}_i}{\partial \hat{\rho}}\right) + t\sum_{i=0}^{m+1} t^i \hat{P}_{i-1}\right] + \tilde{\chi}''(\rho)t\sum_{i=0}^{m} t^i \hat{P}_i.$$

It suffices to show that the desired bounds are satisfied by each of the terms $\boldsymbol{R}_m^1$, $\boldsymbol{R}_m^2$, R_m^1, and R_m^2. The bounds on $\|\boldsymbol{R}_m^1\|_s$ and $\|R_m^1\|_s$, $s \geq 0$, follow from the expressions for the residuals just computed and Theorem 4.5.

We next bound $\|\boldsymbol{R}_m^2\|_s$ and $\|R_m^2\|_s$, $s \geq 0$. Using the expressions for $\hat{\boldsymbol{\Phi}}_i$ and $\hat{P}_i$ given in Theorem 3.3, we can write $\boldsymbol{R}_m^2$ and R_m^2 as a sum of terms all with a common factor of $e^{-c\hat{\rho}}$. Now, because of the presence of the factors $\tilde{\chi}'$ and $\tilde{\chi}''$ in the definitions of $\boldsymbol{R}_m^2$ and R_m^2, each of these terms vanishes for $\rho \leq \rho_0/3$. On the region where $\rho \geq \rho_0/3$

$$e^{-c\hat{\rho}} \leq K_j\left(\frac{c\rho_0}{3}\right)^{-j} t^j =: C_j t^j$$

with $K_j := \max_{x\geq 0} x^j e^{-x} < \infty$, for *any* desired power j. Using this result, referring to the expressions given in Theorem 3.3, and applying Corollary 4.4, it is not difficult to show that for any j and suitable C

$$\|\boldsymbol{R}_m^2\|_s \leq Ct^j\|g\|_{m+s+1/2}, \qquad \|R_m^2\|_s \leq Ct^j\|g\|_{m+s-1/2}. \tag{4.11}$$

Finally, we establish the first estimate when $s = -1$. First we note that

$$\hat{\mathcal{A}}_1^j = \frac{1}{j!}\hat{\mathcal{A}}_1^j + t\hat{\rho}\hat{\mathcal{A}}_1^{j+1}.$$

Substituting this and analogous expressions for $\hat{\mathcal{A}}_2^j$, $\hat{\mathcal{A}}_3^j$, $\hat{\mathcal{A}}_4^j$, and $\hat{\boldsymbol{A}}_6^j$ in (4.10) we get

$$\hat{\boldsymbol{R}}_m^1 = \hat{\boldsymbol{R}}_m^{11} + \hat{\boldsymbol{R}}_m^{12} + \hat{\boldsymbol{R}}_m^{13},$$

where

$$\hat{\boldsymbol{R}}_m^{11} = t^{m+1}\left[\sum_{j=0}^{m}\frac{\hat{\rho}^j}{j!}\left(\hat{\mathcal{A}}_1^j\frac{\partial^2\hat{\boldsymbol{\Phi}}_{m-j}}{\partial\hat{\rho}\partial\theta} + \hat{\mathcal{A}}_2^j\frac{\partial\hat{\boldsymbol{\Phi}}_{m-j}}{\partial\hat{\rho}} + \hat{\mathcal{A}}_3^j\frac{\partial^2\hat{\boldsymbol{\Phi}}_{m-j-1}}{\partial\theta^2} + \hat{\mathcal{A}}_4^j\frac{\partial\hat{\boldsymbol{\Phi}}_{m-j-1}}{\partial\theta} + \hat{\boldsymbol{A}}_6^j\frac{\partial\hat{P}_{m-j-1}}{\partial\theta}\right)\right],$$

$$\hat{\boldsymbol{R}}_m^{12} = t^{m+2}\hat{\mathcal{A}}_3\frac{\partial^2\hat{\boldsymbol{\Phi}}_m}{\partial\theta^2},$$

and

$$\hat{\boldsymbol{R}}_m^{13} = t^{m+2}\left[\sum_{j=0}^{m}\hat{\rho}^{j+1}\left(\hat{\mathcal{A}}_1^{j+1}\frac{\partial^2\hat{\boldsymbol{\Phi}}_{m-j}}{\partial\hat{\rho}\partial\theta} + \hat{\mathcal{A}}_2^{j+1}\frac{\partial\hat{\boldsymbol{\Phi}}_{m-j}}{\partial\hat{\rho}} + \hat{\mathcal{A}}_3^{j+1}\frac{\partial^2\hat{\boldsymbol{\Phi}}_{m-j-1}}{\partial\theta^2} + \hat{\mathcal{A}}_4^{j+1}\frac{\partial\hat{\boldsymbol{\Phi}}_{m-j-1}}{\partial\theta} + \hat{\boldsymbol{A}}_6^{j+1}\frac{\partial\hat{P}_{m-j-1}}{\partial\theta}\right) + \hat{\mathcal{A}}_4\frac{\partial\hat{\boldsymbol{\Phi}}_m}{\partial\theta}\right]. \tag{4.12}$$

By (3.16)

$$\hat{\boldsymbol{R}}_m^{11} = -t^{m+1}\left(\hat{\mathcal{A}}_0 \frac{\partial^2 \hat{\boldsymbol{\Phi}}_{m+1}}{\partial \hat{\rho}^2} + \hat{\boldsymbol{A}}_5 \frac{\partial \hat{P}_{m+1}}{\partial \hat{\rho}}\right),$$

or

$$\tilde{\boldsymbol{R}}_m^{11} = -t^{m+2}\left(\hat{\mathcal{A}}_0 t \frac{\partial^2 \tilde{\boldsymbol{\Phi}}_{m+1}}{\partial \rho^2} + \hat{\boldsymbol{A}}_5 \frac{\partial \tilde{P}_{m+1}}{\partial \rho}\right).$$

Therefore, for any $\boldsymbol{\psi} \in \mathring{H}^1(\Omega)$,

(4.13)

$$\begin{aligned}(\chi \boldsymbol{R}_m^{11}, \boldsymbol{\psi}) &= -t^{m+2}\int_0^L\int_0^{\rho_0} \tilde{\chi}(\rho)\left(\hat{\mathcal{A}}_0 t \frac{\partial^2 \tilde{\boldsymbol{\Phi}}_{m+1}}{\partial \rho^2} + \hat{\boldsymbol{A}}_5 \frac{\partial \tilde{P}_{m+1}}{\partial \rho}\right)\tilde{\boldsymbol{\psi}}(\rho,\theta)[1-\kappa(\theta)\rho]\,d\rho d\theta \\ &= t^{m+2}\int_0^L\int_0^{\rho_0}\left(\hat{\mathcal{A}}_0 t \frac{\partial \tilde{\boldsymbol{\Phi}}_{m+1}}{\partial \rho} + \hat{\boldsymbol{A}}_5 \tilde{P}_{m+1}\right)\frac{\partial}{\partial \rho}\left\{\tilde{\chi}(\rho)\tilde{\boldsymbol{\psi}}(\rho,\theta)[1-\kappa(\theta)\rho]\right\} d\rho d\theta.\end{aligned}$$

Applying the Schwarz inequality and Theorem 4.5 gives

$$(\chi \boldsymbol{R}_m^{11}, \boldsymbol{\psi}) \le C t^{m+5/2}\|g\|_{m+1/2}\|\boldsymbol{\psi}\|_1, \tag{4.14}$$

or, since $\boldsymbol{\psi}$ was arbitrary,

$$\|\chi \boldsymbol{R}_m^{11}\|_{-1} \le C t^{m+5/2}\|g\|_{m+1/2}.$$

Similarly,

$$\begin{aligned}(\chi \boldsymbol{R}_m^{12}, \boldsymbol{\psi}) &= t^{m+2}\int_0^L\int_0^{\rho_0} \tilde{\chi}(\rho)\hat{\mathcal{A}}_3 \frac{\partial^2 \tilde{\boldsymbol{\Phi}}_m}{\partial \theta^2}\tilde{\boldsymbol{\psi}}(\rho,\theta)[1-\kappa(\theta)\rho]\,d\rho d\theta \\ &= -t^{m+2}\int_0^L\int_0^{\rho_0} \tilde{\chi}(\rho)\hat{\mathcal{A}}_3 \frac{\partial \tilde{\boldsymbol{\Phi}}_m}{\partial \theta}\frac{\partial}{\partial \theta}\left\{\tilde{\boldsymbol{\psi}}(\rho,\theta)[1-\kappa(\theta)\rho]\right\} d\rho d\theta,\end{aligned}$$

whence

$$\|\chi \boldsymbol{R}_m^{12}\|_{-1} \le C t^{m+5/2}\|g\|_{m+1/2}.$$

Finally, applying Theorem 4.5 directly to (4.12) and (4.11), respectively, we get

$$\|\chi \boldsymbol{R}_m^{13}\|_{-1} \le C\|\boldsymbol{R}_m^{13}\|_0 \le C t^{m+5/2}\|g\|_{m+1/2},$$

and

$$\|\boldsymbol{R}_m^2\|_{-1} \le \|\boldsymbol{R}_m^2\|_0 \le C t^{m+5/2}\|g\|_{m+1/2}.$$

Since $\boldsymbol{R}_m = \chi \boldsymbol{R}_m^{11} + \chi \boldsymbol{R}_m^{12} + \chi \boldsymbol{R}_m^{13} + \boldsymbol{R}_m^2$, the last three equations imply

$$\|\boldsymbol{R}_m\|_{-1} \le C t^{m+5/2}\|g\|_{m+1/2},$$

as desired. □

5. Error estimates. Let

$$\begin{aligned}\boldsymbol{\phi}_n^E &= \boldsymbol{\phi} - \boldsymbol{\phi}_n^I - \boldsymbol{\phi}_{n-2}^B \\ &= \boldsymbol{\phi} - [\boldsymbol{\phi}_0 + t\boldsymbol{\phi}_1 + \cdots + t^n\boldsymbol{\phi}_n + \chi(t^2\boldsymbol{\Phi}_0 + t^3\boldsymbol{\Phi}_1 + \cdots + t^n\boldsymbol{\Phi}_{n-2})], \\ p_n^E &= p - p_n^I - p_{n-2}^B \\ &= p - [p_0 + tp_1 + \cdots + t^n p_n + \chi(tP_0 + t^2P_1 + \cdots + t^{n-1}P_{n-2})].\end{aligned}$$

Thus $\boldsymbol{\phi}_n^E$ and p_n^E denote the errors in the asymptotic expansions up to order roughly n. Since $\boldsymbol{\phi}_1$ and p_1 vanish,

$$\boldsymbol{\phi}_0^E = \boldsymbol{\phi}_1^E = \boldsymbol{\phi} - \boldsymbol{\phi}_0, \qquad p_0^E = p_1^E = p - p_0.$$

In this section we derive rigorous error bounds for $\boldsymbol{\phi}_n^E$ and p_n^E. In Theorem 5.1 we bound the error in $\boldsymbol{H}^1(\Omega)\times L^2(\Omega)/\mathbb{R}$ and in Theorem 5.2 we bound the error in higher order Sobolev norms.

THEOREM 5.1 (ERROR ESTIMATES FOR $\boldsymbol{\phi}$ AND p IN ENERGY NORM). *There exists a constant C independent of t such that*

$$\|\boldsymbol{\phi}_1^E\|_1 + \|p_1^E\|_{0/\mathbb{R}} + t\|\operatorname{\mathbf{grad}} p_1^E\|_0 \le Ct^{3/2}\|g\|_{-1/2}$$

and for $n \ge 2$

$$\|\boldsymbol{\phi}_n^E\|_1 + \|p_n^E\|_{0/\mathbb{R}} + t\|\operatorname{\mathbf{grad}} p_n^E\|_0 \le C(t^{n+1/2}\|g\|_{n-3/2} + t^{n+3/2}\|g\|_{n-1/2}).$$

Proof. It follows easily from (2.4), (2.5), (2.7), Theorem 4.6, and (4.7)–(4.9) that $(\boldsymbol{\phi}_n^E, p_n^E)$ satisfy the partial differential equations

$$-\operatorname{\mathbf{div}} C\,\mathcal{E}(\boldsymbol{\phi}_n^E) - \operatorname{\mathbf{curl}} p_n^E = -\boldsymbol{R}_{n-2}, \tag{5.1}$$

$$-\operatorname{rot}\boldsymbol{\phi}_n^E + t^2\,\Delta p_n^E = -R_{n-2} \pmod{\mathbb{R}}, \tag{5.2}$$

and the boundary conditions

$$\boldsymbol{\phi}_n^E = 0, \qquad \frac{\partial p_n^E}{\partial n} = -t^{n-1}\frac{\partial p_{n-1}}{\partial n} - t^n\frac{\partial p_n}{\partial n}. \tag{5.3}$$

Writing these equations variationally, we get for all $\boldsymbol{\psi} \in \mathring{\boldsymbol{H}}^1(\Omega)$ and $q \in L^2(\Omega)$ with mean value zero,

$$\begin{aligned}&(C\,\mathcal{E}(\boldsymbol{\phi}_n^E), \mathcal{E}(\boldsymbol{\psi})) - (\operatorname{\mathbf{curl}} p_n^E, \boldsymbol{\psi}) = -(\boldsymbol{R}_{n-2}, \boldsymbol{\psi}), \\ &(\boldsymbol{\phi}_n^E, \operatorname{\mathbf{curl}} q) + t^2(\operatorname{\mathbf{grad}} p_n^E, \operatorname{\mathbf{grad}} q) = (R_{n-2}, q) - t^{n+1}\langle \partial p_{n-1}/\partial n + t\partial p_n/\partial n, q\rangle.\end{aligned} \tag{5.4}$$

Now let

$$\bar{p}_n^E = p_n^E - \frac{1}{\operatorname{meas}\Omega}\int_\Omega p_n^E\,dx$$

denote the difference between p_n^E and its mean value. Choosing $\boldsymbol{\psi} = \boldsymbol{\phi}_n^E$ and $q = \bar{p}_n^E$ and adding the equations, we obtain

$$\begin{aligned}&(C\,\mathcal{E}(\boldsymbol{\phi}_n^E), \mathcal{E}(\boldsymbol{\phi}_n^E)) + t^2(\operatorname{\mathbf{grad}} p_n^E, \operatorname{\mathbf{grad}} p_n^E) \\ &\qquad = -(\boldsymbol{R}_{n-2}, \boldsymbol{\phi}_n^E) + (R_{n-2}, \bar{p}_n^E) - t^{n+1}\langle \partial p_{n-1}/\partial n + t\partial p_n/\partial n, \bar{p}_n^E\rangle.\end{aligned}$$

Applying Korn's inequality and standard estimates, we thus obtain

$$\|\boldsymbol{\phi}_n^E\|_1^2 + t^2\|\operatorname{\mathbf{grad}} p_n^E\|_0^2 \le C(\|\boldsymbol{R}_{n-2}\|_{-1}\|\boldsymbol{\phi}_n^E\|_1 + \|R_{n-2}\|_0\|p_n^E\|_{0/\mathbb{R}} \cdot + t^{n+1}(|\partial p_{n-1}/\partial n|_0 + t|\partial p_n/\partial n|_0)|\bar{p}_n^E|_0).$$

Now

$$|\bar{p}_n^E|_0 \le C\|\bar{p}_n^E\|_0^{1/2}\|\bar{p}_n^E\|_1^{1/2} \le C(t^{-1/2}\|p_n^E\|_{0/\mathbb{R}} + t^{1/2}\|\operatorname{\mathbf{grad}} p_n^E\|_0),$$

so the last term in the previous estimate may be bounded by

$$Ct^{n+1/2}(|\partial p_{n-1}/\partial n|_0 + t|\partial p_{n-1}/\partial n|_0)(\|p_n^E\|_{0/\mathbb{R}} + t\|\operatorname{\mathbf{grad}} p_n^E\|_0).$$

Now choose $\boldsymbol{\psi} \in \mathring{\boldsymbol{H}}^1(\Omega)$ satisfying

$$\operatorname{rot}\boldsymbol{\psi} = \bar{p}_n^E, \qquad \|\boldsymbol{\psi}\|_1 \le C\|p_n^E\|_{0/\mathbb{R}}.$$

(The existence of $\boldsymbol{\psi}$ follows from Lemma 4.2.) From the first variational equation, we obtain

$$\begin{aligned}\|p_n^E\|_{0/\mathbb{R}}^2 &= (p_n^E, \bar{p}_n^E)\\ &= (\boldsymbol{C}\,\boldsymbol{\mathcal{E}}(\boldsymbol{\phi}_n^E), \boldsymbol{\mathcal{E}}(\boldsymbol{\psi})) + (\boldsymbol{R}_{n-2}, \boldsymbol{\psi})\\ &\le C\|\boldsymbol{\psi}\|_1(\|\boldsymbol{\phi}_n^E\|_1 + \|\boldsymbol{R}_{n-2}\|_{-1}),\end{aligned}$$

and so

$$\|p_n^E\|_{0/\mathbb{R}} \le C(\|\boldsymbol{\phi}_n^E\|_1 + \|\boldsymbol{R}_{n-2}\|_{-1}).$$

Combining all these results and using the arithmetic-geometric mean inequality, we obtain

$$\begin{aligned}&\|\boldsymbol{\phi}_n^E\|_1 + \|p_n^E\|_{0/\mathbb{R}} + t\|\operatorname{\mathbf{grad}} p_n^E\|_0\\ &\qquad \le C\left[\|\boldsymbol{R}_{n-2}\|_{-1} + \|R_{n-2}\|_0 + t^{n+1/2}|\partial p_{n-1}/\partial n|_0 + t^{n+3/2}|\partial p_n/\partial n|_0\right].\end{aligned}$$

Note that if $n = 1$, the right-hand side reduces to $Ct^{3/2}|\partial p_0/\partial n|_0$. The theorem follows immediately from this estimate, Corollary 4.4, and Theorem 4.7. □

We now turn to the derivation of error estimates in higher norms.

THEOREM 5.2 (ERROR ESTIMATES FOR $\boldsymbol{\phi}$ AND p IN HIGHER NORMS). *Let $s \ge 2$ be an integer. Then*

$$\|\boldsymbol{\phi}_1^E\|_s + t\|p_1^E\|_{s/\mathbb{R}} \le C(t^{5/2-s}\|g\|_{-1/2} + t\|g\|_{s-2})$$

and for $n \ge 2$

$$\|\boldsymbol{\phi}_n^E\|_s + t\|p_n^E\|_{s/\mathbb{R}} \le C(t^{n+3/2-s}\|g\|_{n-3/2} + t^{n+1}\|g\|_{n+s-2}).$$

Proof. By standard regularity results for the Dirichlet problem for plane elasticity and (5.1),

$$\|\boldsymbol{\phi}_n^E\|_s \le C\|\operatorname{\mathbf{div}} \boldsymbol{C}\,\boldsymbol{\mathcal{E}}(\boldsymbol{\phi}_n^E)\|_{s-2} \le C(\|\operatorname{\mathbf{grad}} p_n^E\|_{s-2} + \|\boldsymbol{R}_{n-2}\|_{s-2}). \tag{5.5}$$

Using regularity for the Neumann problem for the Laplacian and (5.2) and (5.3), we similarly obtain

$$\|p_n^E\|_{s/\mathbb{R}} \le C\left(\|\Delta p_n^E\|_{s-2/\mathbb{R}} + \left|\frac{\partial p_n^E}{\partial n}\right|_{s-3/2}\right)$$
$$\le C\left(t^{-2}\|\operatorname{rot}\boldsymbol{\phi}_n^E\|_{s-2} + t^{-2}\|R_{n-2}\|_{s-2} + t^{n-1}\left|\frac{\partial p_{n-1}}{\partial n}\right|_{s-3/2} + t^n\left|\frac{\partial p_n}{\partial n}\right|_{s-3/2}\right).$$

Combining these results and using Corollary 4.4 and Theorem 4.7, we get for $n \ge 1$, $s \ge 2$,

$$\|\boldsymbol{\phi}_n^E\|_s + t\|p_n^E\|_{s/\mathbb{R}} \le C\left(\|\mathbf{grad}\, p_n^E\|_{s-2} + \|\boldsymbol{R}_{n-2}\|_{s-2} + t^{-1}\|\operatorname{rot}\boldsymbol{\phi}_n^E\|_{s-2}\right.$$
$$\left. + t^{-1}\|R_{n-2}\|_{s-2} + t^n\left|\frac{\partial p_{n-1}}{\partial n}\right|_{s-3/2} + t^{n+1}\left|\frac{\partial p_n}{\partial n}\right|_{s-3/2}\right)$$
$$\le C(\|p_n^E\|_{s-1/\mathbb{R}} + t^{-1}\|\boldsymbol{\phi}_n^E\|_{s-1} + t^{n+3/2-s}\|g\|_{n-3/2}$$
$$+ t^{n+1/2}\|g\|_{n+s-5/2} + t^n\|g\|_{n+s-3} + t^{n+1}\|g\|_{n+s-2}).$$

Since $\boldsymbol{R}_{-1}$, R_{-1}, and p_1 vanish, for $n = 1$ we can simplify this result to

$$\|\boldsymbol{\phi}_1^E\|_s + t\|p_1^E\|_{s/\mathbb{R}} \le C(\|p_1^E\|_{s-1/\mathbb{R}} + t^{-1}\|\boldsymbol{\phi}_1^E\|_{s-1} + t\|g\|_{s-2}).$$

Thus

$$\|\boldsymbol{\phi}_n^E\|_s + t\|p_n^E\|_{s/\mathbb{R}}$$
$$\le \begin{cases} C(\|p_1^E\|_{s-1/\mathbb{R}} + t^{-1}\|\boldsymbol{\phi}_1^E\|_{s-1} + t\|g\|_{s-2}), & n = 1, \\ C(\|p_n^E\|_{s-1/\mathbb{R}} + t^{-1}\|\boldsymbol{\phi}_n^E\|_{s-1} + t^{n+3/2-s}\|g\|_{n-3/2} + t^{n+1}\|g\|_{n+s-2}), & n \ge 2. \end{cases}$$

For $s = 2$, the theorem follows from this relation and Theorem 5.1. We can complete the proof using this relation and a simple induction on s. □

As a consequence of Theorems 5.1, 5.2, and 4.3, we easily obtain bounds on $\boldsymbol{\phi}$ and p.

THEOREM 5.3 (BOUNDS ON $\boldsymbol{\phi}$ AND p).

$$\|\boldsymbol{\phi}\|_s \le C(t^{5/2-s}\|g\|_{-1/2} + t\|g\|_{s-2} + |||g|||_{s-3}), \qquad s = 1, 2, \cdots,$$
$$\|p\|_{s/\mathbb{R}} \le C(t^{3/2-s}\|g\|_{-1/2} + |||g|||_{s-2}), \qquad s \in \mathbb{N}.$$

Proof. From Theorems 5.1 and 5.2 we have

$$\|p_1^E\|_{s/\mathbb{R}} \le C(t^{3/2-s}\|g\|_{-1/2} + |||g|||_{s-2}), \qquad s \in \mathbb{N}.$$

By Theorem 4.3,

$$\|p_0\|_{s/\mathbb{R}} \le C|||g|||_{s-2}.$$

Applying the triangle inequality, we obtain

$$\|p\|_{s/\mathbb{R}} \le \|p_1^E\|_{s/\mathbb{R}} + \|p_0\|_{s/\mathbb{R}} \le C(t^{3/2-s}\|g\|_{-1/2} + |||g|||_{s-2}).$$

A similar argument gives the estimate on ϕ. □

We may use the interpolation property of the Sobolev norms to obtain bounds on $\|\phi_n^E\|_s$ and $\|p_n^E\|_s$ for noninteger s similar to those given in Theorems 5.1 and 5.2 for integer s. In particular we have

$$\begin{aligned}\|\phi_1^E\|_{5/2} &\le C(\|\phi_1^E\|_2\|\phi_1^E\|_3)^{1/2}\\ &\le C[(t^{1/2}\|g\|_{-1/2}+t\|g\|_0)(t^{-1/2}\|g\|_{-1/2}+t\|g\|_1)]^{1/2}\\ &\le C(\|g\|_{-1/2}+t^{3/2}\|g\|_1)\end{aligned}$$

and, similarly,

$$\begin{aligned}\|p_1^E\|_{3/2/\mathbf{R}} &\le C(\|p_1^E\|_{1/\mathbf{R}}\|p_1^E\|_{2/\mathbf{R}})^{1/2}\\ &\le C[t^{1/2}\|g\|_{-1/2}(t^{-1/2}\|g\|_{-1/2}+\|g\|_0)]^{1/2}\\ &\le C(\|g\|_{-1/2}+t^{1/2}\|g\|_0).\end{aligned}$$

Combining with Theorem 4.3 as above, we get

$$\|\phi\|_{5/2} \le C(\|g\|_{-1/2}+t^{3/2}\|g\|_1), \tag{5.6}$$

$$\|p\|_{3/2/\mathbf{R}} \le C(\|g\|_{-1/2}+t^{1/2}\|g\|_0). \tag{5.7}$$

In general, however, higher norms of ϕ and p do not remain bounded as $t \to 0$.

Thus far our estimates have all been in the L^2-based Sobolev spaces H^s. In closing this section, we note that our asymptotic expansions and error estimates can be used to study the dependence of the solution on t in many other function spaces as well, for example in the L^p-based Sobolev spaces W_p^s or the Hölder spaces $C^{m,\alpha}$. To determine the behavior of the norm $\|\phi\|_{W_\infty^s}$ with respect to t, for example, we may write $\phi = \phi_n^E + \phi_n^I + \phi_{n-2}^B$. Now, assuming g is sufficiently smooth, $\|\phi_n^E\|_{n+3/2}$ is bounded uniformly in t. Hence, if n is sufficiently large ($n > s - \frac{1}{2}$ in this case), then the Sobolev Embedding Theorem implies that $\|\phi_n^E\|_{W_\infty^s}$ is bounded uniformly. Each of the interior expansion functions is bounded in all the H^s spaces, so $\|\phi_n^I\|_{W_\infty^s}$ is also bounded uniformly. Thus the behavior of ϕ is determined by that of $\phi_{n-2}^B = \chi(t^2\Phi_0+t^3\Phi_1+\cdots+t^n\Phi_{n-2})$. Since we have quite explicit expressions for the boundary correctors (Theorem 3.3), it is not difficult to determine the behavior of ϕ_{n-2}^B. We see that $\|\phi_{n-2}^B\|_{W_\infty^s} = O(t^{2-s})$. Thus

$$\|\phi\|_{W_\infty^s} = O(t^{\min(2-s,0)}).$$

Estimates of other quantities, including the errors in the partial sums of the asymptotic expansions can be derived similarly. With a little effort we can get a bound which indicates explicitly the dependence of the norm on the load function g as well. However, we do not expect that the required regularity on g in these estimates (and in some of the previous ones as well) is optimal.

6. Asymptotic expansion of the transverse displacement and shear. In the previous sections we obtained and justified an asymptotic expansion for the rotation variable ϕ. We now turn to the other primitive variable, ω, and obtain an expansion for it. In contrast to ϕ, we will see that ω has no boundary layer.

Define the auxiliary variable $v = \omega - t^2 r$. Clearly $v = 0$ on $\partial\Omega$ and, from (2.6), $\Delta v = \operatorname{div} \phi$. Then, taking the divergence of (2.4) and substituting (2.3), we easily compute that $D\Delta^2 v = D\Delta \operatorname{div} \phi = g$. Next, note that $\mathbf{grad}\, v = \mathbf{grad}\, w - t^2\, \mathbf{grad}\, r = \phi + t^2\, \mathbf{curl}\, p$. Since ϕ vanishes on $\partial\Omega$,

$$\frac{\partial v}{\partial n} = -t^2 \frac{\partial p}{\partial s} \quad \text{on } \partial\Omega.$$

Thus v is completely characterized as the solution of a certain Dirichlet problem for the biharmonic operator, and it is easy to see how to expand it in powers of t. For $i \in \mathbb{N}$, define v_i by the biharmonic problem

$$D\Delta^2 v_i = \begin{cases} g, & i = 0, \\ 0, & i \geq 1, \end{cases} \quad \text{in } \Omega,$$

$$v_i = 0, \quad \partial v_i/\partial n = \begin{cases} 0, & i = 0, 1, \\ -\partial p_0/\partial s, & i = 2 \\ -\partial p_{i-2}/\partial s - \partial P_{i-3}/\partial s, & i \geq 3, \end{cases} \quad \text{on } \partial\Omega.$$

The coefficients in the asymptotic expansion of ω are then given by

$$\omega_i = \begin{cases} v_i, & i \neq 2, \\ v_2 + r, & i = 2. \end{cases}$$

Note that ω_0 satisfies the boundary value problem

$$D\Delta^2 \omega_0 = g \quad \text{in } \Omega, \qquad \omega_0 = \frac{\partial \omega_0}{\partial n} = 0 \quad \text{on } \partial\Omega.$$

It is useful to express the first terms of the expansions for w and ϕ in terms of ω_0. First of all, there is a simple relation between ω_0 and ϕ_0.

THEOREM 6.1.

$$\phi_0 = \mathbf{grad}\, \omega_0.$$

Proof. From (3.10) and (3.9), it follows that $\phi_0 = \mathbf{grad}\, \mu$ for some $\mu \in \mathring{H}^2(\Omega)$. Inserting in (3.7) and taking the divergence gives

$$D\Delta^2 \mu = -\Delta r = g.$$

Comparing with the defining equations for ω_0, we see that $\mu = \omega_0$. □

Clearly $\omega_1 = v_1 = 0$ and $\omega_2 = v_2 + r$, where

$$\Delta^2 v_2 = 0 \quad \text{in } \Omega, \qquad v_2 = 0, \quad \partial v_2/\partial n = -\partial p_0/\partial s \quad \text{on } \partial\Omega, \tag{6.1}$$

$$\Delta r = -g \quad \text{in } \Omega, \qquad r = 0 \quad \text{on } \partial\Omega. \tag{6.2}$$

Now, from (3.7),

$$\frac{\partial p_0}{\partial s} = \mathbf{div}\, C\, \mathcal{E}(\phi_0) \cdot \boldsymbol{n} + \frac{\partial r}{\partial n} = D\frac{\partial \Delta \omega_0}{\partial n} + \frac{\partial r}{\partial n}. \tag{6.3}$$

$$\frac{\partial p_0}{\partial n} = -\mathbf{div}\, C\, \mathcal{E}(\phi_0) \cdot \boldsymbol{s} = -D\frac{\partial \Delta \omega_0}{\partial s}. \tag{6.4}$$

Using (6.3) in (6.1) and combining with (6.2), we get

$$\Delta^2 \omega_2 = -\Delta g \quad \text{on } \Omega, \qquad \omega_2 = 0, \quad \frac{\partial \omega_2}{\partial n} = -D\frac{\partial \Delta \omega_0}{\partial n},$$

which is a biharmonic problem for ω_2. From the definitions, $\omega_3 = v_3$ is a biharmonic function vanishing on $\partial\Omega$ with $\partial\omega_3/\partial n = -\partial P_0/\partial s$ (since $p_1 = 0$). Using (3.20) and (6.4) to simplify the latter boundary condition gives the following biharmonic problem for ω_3:

$$\Delta^2 \omega_3 = 0 \quad \text{in } \Omega, \qquad \omega_3 = 0, \quad \frac{\partial \omega_3}{\partial n} = -\frac{D}{c}\frac{\partial^2}{\partial s^2}\Delta\omega_0 \quad \text{on } \partial\Omega.$$

Turning to the expansion for $\boldsymbol{\phi}$, the expression for $\boldsymbol{\Phi}_0$ in (3.20) becomes, in light of (6.4),

$$\hat{\boldsymbol{\Phi}}_0(\hat{\rho}, \theta) = -D\frac{\partial \widehat{\Delta \omega_0}}{\partial s}(0, \theta) e^{-c\hat{\rho}} \boldsymbol{s}. \tag{6.5}$$

To determine $\boldsymbol{\phi}_2$, we note from (3.10) that $\operatorname{rot} \boldsymbol{\phi}_2$ is constant. Since

$$\int_\Omega \operatorname{rot} \boldsymbol{\phi}_2 = -\int_{\partial\Omega} \boldsymbol{\phi}_2 \cdot \boldsymbol{s} = \int_{\partial\Omega} \boldsymbol{\Phi}_0 \cdot \boldsymbol{s} = 0,$$

$\operatorname{rot} \boldsymbol{\phi}_2 = 0$ and $\boldsymbol{\phi}_2 = \mathbf{grad}\, \psi$ for some function ψ. Substituting in (3.7) and taking the divergence shows that ψ is biharmonic. Then the boundary conditions $\mathbf{grad}\, \psi = -\boldsymbol{\Phi}_0$ on $\partial\Omega$ determine ψ modulo $\mathbb{R}$ and $\boldsymbol{\phi}_2$ completely. In light of (6.5), the boundary conditions on ψ become

$$\psi = D\Delta\omega_0 \pmod{\mathbb{R}}, \quad \partial\psi/\partial n = 0 \quad \text{on } \partial\Omega.$$

We now obtain a priori estimates for the ω_i and error estimates for the finite sums of the expansion.

THEOREM 6.2 (A PRIORI ESTIMATES FOR THE ω_i). *Let $i \in \mathbb{N}$, $s \geq 2$. Then*

$$\|\omega_i\|_s \leq C |||g|||_{s+i-4}.$$

Proof. This follows easily from regularity for the biharmonic equation, Corollary 4.4, and (4.5). □

Let $\omega_n^E = \omega - \sum_{i=0}^n t^i \omega_i$ denote the error in the partial sums of the asymptotic expansion

$$\omega \sim \sum_{i=0}^{\infty} t^i \omega_i.$$

The next theorem bounds the error in expansion. Note that the order of the error is the same in all Sobolev norms, reflecting the fact that ω does not involve a boundary layer.

THEOREM 6.3 (ERROR ESTIMATES FOR ω). *For $n = 1, 2, \cdots$ and $s = 1, 2, \cdots$*

$$\|\omega_n^E\|_s \leq C(t^{n+1}\|g\|_{n+s-3} + t^{n+s+1}\|g\|_{n+2s-3}).$$

Proof. Set $v_n^E = v - \sum_{i=0}^n t^i v_i$. Note that $\omega_n^E = v_n^E$ for $n > 1$, and $\omega_1^E = v_1^E + t^2 r$, so it suffices to prove the theorem with ω_n^E replaced by v_n^E.

Now

$$D\,\Delta^2 v_n^E = 0 \quad \text{in } \Omega,$$
$$v_n^E = 0 \quad \text{on } \partial\Omega,$$

and

$$\begin{aligned}\frac{\partial v_n^E}{\partial n} &= -t^2 \frac{\partial}{\partial s}(p - p_{n-2}^I - p_{n-3}^B) \\ &= -t^2 \frac{\partial}{\partial s}\left(p_{n+s-1}^E + \sum_{i=n-1}^{n+s-1} t^i p_i + \sum_{i=n-1}^{n+s-2} t^i P_{i-1}\right).\end{aligned}$$

Thus, using regularity results for the biharmonic problem,

$$\begin{aligned}\|v_n^E\|_s &\le C \left|\frac{\partial v_n^E}{\partial n}\right|_{s-3/2} \\ &\le Ct^2 \left(\|p_{n+s-1}^E\|_{s/\mathbf{R}} + \sum_{i=n-1}^{n+s-1} t^i \|p_i\|_{s/\mathbf{R}} + \sum_{i=n-1}^{n+s-2} t^i |P_{i-1}|_{s-1/2}\right).\end{aligned}$$

Applying Theorem 5.1 or 5.2, Theorem 4.3, and Corollary 4.4, we get

$$\begin{aligned}\|v_n^E\|_s &\le Ct^2 \left(t^{n-1/2}\|g\|_{n+s-5/2} + t^{n+s-1}\|g\|_{n+2s-3} + \sum_{i=n-1}^{n+s-1} t^i \|g\|_{i+s-2}\right) \\ &\le C(t^{n+1}\|g\|_{n+s-3} + t^{n+s+1}\|g\|_{n+2s-3}),\end{aligned}$$

as desired. □

Using a similar argument, we can also obtain regularity estimates for ω in $H^s(\Omega)$ uniform with respect to t.

THEOREM 6.4. *For $s = 2, 3, \cdots$ there exists a constant C independent of t such that*

$$\|\omega\|_s \le C(\|g\|_{s-4} + t^2\|g\|_{s-2}), \quad s = 2, 3,$$
$$\|\omega\|_s \le C(\|g\|_{s-4} + t^s\|g\|_{2s-4}), \quad s \ge 4.$$

Proof. Using standard regularity results for the biharmonic problem and (4.5), we get

$$\|\omega\|_s \le \|v\|_s + t^2\|r\|_s \le C(\|g\|_{s-4} + t^2|\partial p/\partial s|_{s-3/2} + t^2\|g\|_{s-2}). \tag{6.6}$$

When $s \ge 4$, we substitute $p = \sum_{i=0}^{s-2} t^i p_i + \chi \sum_{i=0}^{s-4} t^{i+1} P_i + p_{s-2}^E$, into (6.6) and apply Corollary 4.4 and Theorem 5.2 to estimate the right-hand side, obtaining

$$\|\omega\|_s \le C\left(\|g\|_{s-4} + \sum_{i=0}^{s-2} t^{i+2}\|g\|_{s+i-2} + \sum_{i=0}^{s-4} t^{i+3}\|g\|_{s+i-1} + t^{1/2}\|g\|_{s-7/2}\right),$$

which gives the desired result.

When $s = 2$ or $s = 3$, we substitute $p = p_0 + p_1^E$ into (6.6) and complete the proof with a similar argument. □

Recall that the scaled transverse shear stress is given by $\boldsymbol{\zeta} = t^{-2}(\mathbf{grad}\,\omega - \boldsymbol{\phi})$, which we decomposed as $\mathbf{grad}\,r + \mathbf{curl}\,p$. We can obtain an asymptotic expansion for the shear stress from either of these expressions, in the former case noting a cancellation due to Theorem 6.1. Thus, formally,

$$\begin{aligned}\boldsymbol{\zeta} &\sim (\mathbf{grad}\,\omega_2 - \boldsymbol{\phi}_2) + t(\mathbf{grad}\,\omega_3 - \boldsymbol{\phi}_3) + \cdots - \chi(\boldsymbol{\Phi}_0 + t\boldsymbol{\Phi}_1 + \cdots)\\ &\sim (\mathbf{grad}\,r + \mathbf{curl}\,p_0) + t^2\,\mathbf{curl}\,p_2 + t^3\,\mathbf{curl}\,p_3 + \cdots + \chi(t\,\mathbf{curl}\,P_0 + t^2\,\mathbf{curl}\,P_1 + \cdots).\end{aligned}$$

In light of our previous results, it is straightforward to bound the individual terms in either of these expansions as well as the remainders when the expansions are terminated. Here we content ourselves with determining the regularity of the shear stress vector and its dependence on t.

THEOREM 6.5. *Let $s \geq -1$ be an integer. Then there exists a constant C independent of t such that*

$$\|\boldsymbol{\zeta}\|_s \leq C(t^{1/2-s}\|g\|_{-1/2} + |||g|||_{s-1}).$$

Proof. This follows immediately from Theorem 5.3 and (4.5). □

Similar bounds hold in the noninteger order Sobolev spaces. In particular,

$$\|\boldsymbol{\zeta}\|_{1/2} \leq C(\|g\|_{-1/2} + t^{1/2}\|g\|_0),$$

as follows immediately from (4.5) and (5.7). In general $\|\boldsymbol{\zeta}\|_s$ will blow up as $t \to 0$ if $s > 1/2$. Thus the shear stress evidences a rather strong boundary layer.

7. Hard simply-supported boundary conditions. Two sets of boundary conditions are commonly used with the Reissner–Mindlin equations to model a simply-supported plate. Boundary conditions for a hard simply-supported plate are

$$M_n\boldsymbol{\phi} = 0, \quad \boldsymbol{\phi}\cdot\boldsymbol{s} = 0, \quad w = 0,$$

where $M_n\boldsymbol{\phi} = \boldsymbol{n}^t\mathcal{C}\,\mathcal{E}(\boldsymbol{\phi})\boldsymbol{n}$, or, in (ρ,θ) coordinates,

$$\widetilde{M_n\boldsymbol{\phi}} = D\left(-\frac{\partial\tilde{\boldsymbol{\phi}}}{\partial\rho}\cdot\boldsymbol{n} + \nu\frac{\partial\tilde{\boldsymbol{\phi}}}{\partial\theta}\cdot\boldsymbol{s}\right).$$

(For a soft simply-supported plate the condition $\boldsymbol{\phi}\cdot\boldsymbol{s} = 0$ is replaced by $\boldsymbol{s}^t\mathcal{C}\,\mathcal{E}(\boldsymbol{\phi})\boldsymbol{n} = 0$. Thus, in both cases the lateral edge of the undisplaced plate is not permitted to displace vertically. In the soft case a vertical fiber on the lateral edge is permitted to rotate freely, while in the hard case it may only rotate in the plane normal to the edge. The soft conditions would seem to be easier to realize in practice.)

The boundary layer analysis for the hard simply-supported plate, which we consider in this section, is very similar to that for the clamped plate. The soft simply-supported plate has a significantly stronger boundary layer which will be investigated in a subsequent paper.

The only difference in the asymptotic expansions themselves for the hard simply-supported Reissner–Mindlin plate and the clamped plate is that the boundary conditions for the problems defining the interior expansion functions must be modified. All

the major estimates for the expansion functions and all the error analysis carries over. However, at a few places in the analysis additional terms must be considered. In this section we indicate very briefly these additional considerations.

As in the case of the clamped plate, we use the decomposition of the shear stress vector, given by formula (2.2). We then obtain the reformulation of (2.3)–(2.6), where the boundary conditions (2.7) are replaced by:

$$r = 0, \quad \boldsymbol{\phi} \cdot \boldsymbol{s} = 0, \quad M_n \boldsymbol{\phi} = 0, \quad \partial p / \partial n = 0, \quad w = 0.$$

The forms of the asymptotic expansions for $\boldsymbol{\phi}$ and p are the same as those given in § 3 for the clamped plate, and the interior approximations satisfy the same partial differential equations (3.7), (3.10), but the boundary conditions (3.9) are replaced by

$$\boldsymbol{\phi}_i \cdot \boldsymbol{s} = \begin{cases} 0, & i = 0, 1, \\ -\hat{\boldsymbol{\Phi}}_{i-2} \cdot \boldsymbol{s}, & i \geq 2, \end{cases}$$

and

$$M_n \boldsymbol{\phi}_i = \begin{cases} 0, & i = 0, \\ D(\partial \hat{\boldsymbol{\Phi}}_0 / \partial \hat{\rho}) \cdot \boldsymbol{n}, & i = 1, \\ D[(\partial \hat{\boldsymbol{\Phi}}_{i-1} / \partial \hat{\rho}) \cdot \boldsymbol{n} - \nu (\partial \hat{\boldsymbol{\Phi}}_{i-2} / \partial \theta) \cdot \boldsymbol{s}], & i \geq 2. \end{cases}$$

The boundary correctors are again defined by (3.16)–(3.19). Thus, the analysis in § 3 remains valid. In particular, Theorem 3.3 still holds and the formula for the first boundary corrector is again (3.20). It follows immediately that $M_n \boldsymbol{\phi}_1 = 0$ on $\partial\Omega$ and hence $\boldsymbol{\phi}_1 = 0$, $p_1 = 0$.

To bound the errors in the asymptotic expansions, we need analogues of the results proved in § 4. From the form of the boundary correctors (given in Theorem 3.3), we get immediately that

$$|\boldsymbol{\Phi}_i|_s + t \left| \frac{\partial \boldsymbol{\Phi}_i}{\partial n} \right|_s \leq C \sum_{j=0}^{i} \left| \frac{\partial p_j}{\partial n} \right|_{s+i-j}, \tag{7.1}$$

for all $s \in \mathbb{R}$, $i \in \mathbb{N}$. To estimate the interior expansion functions we use the following analogue of Lemma 4.2.

LEMMA 7.1. *Let $s \in \mathbb{N}$, $f \in H^s(\Omega) \cap \mathring{H}^1(\Omega)$, $g \in H^s(\Omega)/\mathbb{R}$, $k \in H^{s+1/2}(\partial\Omega)$, and $l \in H^{s-1/2}(\partial\Omega)$ be given. Then there exist unique $\boldsymbol{\psi} \in \boldsymbol{H}^{s+1}(\Omega)$, $q \in H^s(\Omega)/\mathbb{R}$ satisfying the partial differential equations*

$$-\operatorname{\mathbf{div}} \boldsymbol{C}\, \boldsymbol{\mathcal{E}}(\boldsymbol{\psi}) - \operatorname{\mathbf{curl}} q = \operatorname{\mathbf{grad}} f, \tag{7.2}$$

$$-\operatorname{rot} \boldsymbol{\psi} = g \pmod{\mathbb{R}}, \tag{7.3}$$

and the boundary conditions

$$\boldsymbol{\psi} \cdot \boldsymbol{s} = k, \qquad M_n \boldsymbol{\psi} = l.$$

Moreover, there exists a constant C depending only on s, E, ν, and Ω such that

$$\|\boldsymbol{\psi}\|_{s+1} + \|q\|_{s/\mathbb{R}} \leq C(\|f\|_s + \|g\|_{s/\mathbb{R}} + |k|_{s+1/2} + |l|_{s-1/2}).$$

Proof. The weak form of the boundary value problem is to find $\boldsymbol{\psi} \in \boldsymbol{H}^1(\Omega)$ such that $\boldsymbol{\psi} \cdot \boldsymbol{s} = k$ on $\partial\Omega$ and $q \in L^2(\Omega)/\mathbb{R}$ satisifying

$$\begin{aligned}(\mathcal{C}\,\mathcal{E}(\boldsymbol{\psi}), \mathcal{E}(\boldsymbol{\mu})) - (q, \operatorname{rot} \boldsymbol{\mu}) &= -(f, \operatorname{div} \boldsymbol{\mu}) + \langle l, \boldsymbol{\mu} \cdot \boldsymbol{n} \rangle, \\ -(\operatorname{rot} \boldsymbol{\psi}, v) &= (g, v),\end{aligned}$$

for all $\boldsymbol{\mu} \in \boldsymbol{H}^1(\Omega)$ such that $\boldsymbol{\mu} \cdot \boldsymbol{s} = 0$ on $\partial\Omega$ and all $v \in L^2(\Omega)$ of mean value zero. Existence and uniqueness of a solution in $\boldsymbol{H}^1(\Omega) \times L^2(\Omega)/\mathbb{R}$ is proved just as for the generalized Stokes equations, e.g., by applying Brezzi's theorem [4]. The estimate for $s = 0$ follows from the same argument. To establish the claimed regularity for $s \geq 1$, we apply the Helmholtz decomposition to $\boldsymbol{\psi}$ to get $\boldsymbol{\psi} = \mathbf{grad}\, z + \mathbf{curl}\, b$, with $z \in \mathring{H}^1(\Omega)$, $b \in H^1(\Omega)/\mathbb{R}$. Now, it suffices to show that

$$\|b\|_{s+2/\mathbb{R}} + \|z\|_{s+2} \leq C(\|f\|_s + \|g\|_{s/\mathbb{R}} + |k|_{s+1/2}), \quad s \in \mathbb{N},$$

since this gives the estimate on $\boldsymbol{\psi}$ immediately, and that on q then follows from (7.2).

From (7.3) we have

$$\Delta b = g \pmod{\mathbb{R}} \quad \text{in } \Omega,$$

with boundary conditions $\partial b/\partial n = k$, so the desired bound on b follows from regularity for the Neumann problem for Laplace's equation. We prove the desired estimate for z by induction on s. The case $s = 0$ follows from the bound on $\|\boldsymbol{\psi}\|_1$ since $z = \operatorname{div} \boldsymbol{\psi}$. Thus we assume that s is a positive integer.

Let $w = D\,\Delta z + f$. Since

$$\begin{aligned}M_n(\mathbf{grad}\, z) &= D\left[\Delta z - (1-\nu)\left(\frac{\partial^2 z}{\partial s^2} + \kappa \frac{\partial z}{\partial n}\right)\right], \\ M_n(\mathbf{curl}\, b) &= D(1-\nu)\left(-\frac{\partial}{\partial s}\frac{\partial b}{\partial n} + \kappa \frac{\partial b}{\partial s}\right),\end{aligned}$$

the boundary conditions for $\boldsymbol{\psi}$ imply that

$$w - f = D\,\Delta z = l + D(1-\nu)(\partial^2 z/\partial s^2 + \kappa \partial z/\partial n + \partial k/\partial s - \kappa \partial b/\partial s) \quad \text{on } \partial\Omega,$$

or, since z and f vanish on $\partial\Omega$,

$$w = l + D(1-\nu)(\kappa \partial z/\partial n + \partial k/\partial s - \kappa \partial b/\partial s) \quad \text{on } \partial\Omega.$$

Now, taking the divergence of equation (7.2) gives

$$-D\,\Delta^2 z = \Delta f \quad \text{in } \Omega.$$

so w is harmonic. Applying regularity for the Dirichlet problem for Laplace's equation then gives

$$\begin{aligned}\|w\|_s &\leq C(|\partial z/\partial n|_{s-1/2} + |\partial k/\partial s|_{s-1/2} + |\partial b/\partial s|_{s-1/2}) \\ &\leq C(\|z\|_{s+1} + |k|_{s+1/2} + \|b\|_{s+1/\mathbb{R}}) \\ &\leq C(\|z\|_{s+1} + |k|_{s+1/2} + \|g\|_{s/\mathbb{R}}).\end{aligned}$$

Finally z satisfies

$$-\Delta z = D^{-1}(f - w) \quad \text{in } \Omega, \qquad z = 0 \quad \text{on } \partial\Omega,$$

so another application of regularity for the Dirichlet problem shows that

$$\|z\|_{s+2} \le C(\|f\|_s + \|w\|_s) \le C(\|f\|_s + \|g\|_{s/\mathbf{R}} + |k|_{s+1/2} + \|z\|_{s+1}),$$

and the proof is completed by induction. □

Using this result and (7.1), it follows that Theorem 4.3 holds also in the hard simply-supported case, and then that Theorem 4.5 also remains valid.

Turning to the finite interior and boundary expansions, Theorem 4.6 and Theorem 4.7 hold as before. However, in order to prove the analogue of Theorem 5.1, we need a slight refinement of the estimate of $\|\boldsymbol{R}_m\|_{-1}$.

THEOREM 7.2. *If $\boldsymbol{\psi} \in \boldsymbol{H}^1(\Omega)$ satisfies $\boldsymbol{\psi} \cdot \boldsymbol{s} = 0$ on $\partial\Omega$ and $m \in \mathbb{N}$, then*

$$\left|(\boldsymbol{R}_m, \boldsymbol{\psi}) - t^{m+3}\left\langle D\frac{\partial \boldsymbol{\Phi}_{m+1}}{\partial n} \cdot \boldsymbol{n}, \boldsymbol{\psi} \cdot \boldsymbol{n}\right\rangle\right| \le Ct^{m+5/2}\|g\|_{m+1/2}\|\boldsymbol{\psi}\|_1.$$

Proof. The proof is very close to that of the H^{-1} estimate in Theorem 4.7. The only difference is that instead of (4.14) we must show that

$$\left|(\chi \boldsymbol{R}_m^{11}, \boldsymbol{\psi}) - t^{m+3}\left\langle D\frac{\partial \boldsymbol{\Phi}_{m+1}}{\partial n} \cdot \boldsymbol{n}, \boldsymbol{\psi} \cdot \boldsymbol{n}\right\rangle\right| \le Ct^{m+5/2}\|g\|_{m+1/2}\|\boldsymbol{\psi}\|_1 \tag{7.4}$$

(which is the same as (4.14) for $\boldsymbol{\psi} \in \mathring{\boldsymbol{H}}^1(\Omega)$). Since $\boldsymbol{\psi}$ does not vanish on the boundary, when we integrate by parts in (4.13) we get a boundary term:

$$\begin{aligned}(\chi \boldsymbol{R}_m^{11}, \boldsymbol{\psi}) &= t^{m+2}\int_0^L\int_0^{\rho_0}\left(\hat{\mathcal{A}}_0 t\frac{\partial \tilde{\boldsymbol{\Phi}}_{m+1}}{\partial \rho} + \hat{\boldsymbol{A}}_5 \tilde{P}_{m+1}\right)\frac{\partial}{\partial \rho}\left\{\tilde{\chi}(\rho)\tilde{\boldsymbol{\psi}}(\rho,\theta)[1-\kappa(\theta)\rho]\right\} d\rho d\theta \\ &\qquad - t^{m+2}\left\langle \mathcal{A}_0 t\frac{\partial \boldsymbol{\Phi}_{m+1}}{\partial n} - \boldsymbol{A}_5 P_{m+1}, \boldsymbol{\psi}\right\rangle.\end{aligned}$$

Now $\boldsymbol{A}_5 \cdot \boldsymbol{\psi} = \boldsymbol{s} \cdot \boldsymbol{\psi} \equiv 0$ and $\mathcal{A}_0\boldsymbol{\psi} = \mathcal{A}_0\boldsymbol{n}(\boldsymbol{\psi} \cdot \boldsymbol{n}) = -D\boldsymbol{n}(\boldsymbol{\psi} \cdot \boldsymbol{n})$, so

$$\left\langle \mathcal{A}_0 t\frac{\partial \boldsymbol{\Phi}_{m+1}}{\partial n} - \boldsymbol{A}_5 P_{m+1}, \boldsymbol{\psi}\right\rangle = -t\left\langle D\frac{\partial \boldsymbol{\Phi}_{m+1}}{\partial n} \cdot \boldsymbol{n}, \boldsymbol{\psi} \cdot \boldsymbol{n}\right\rangle.$$

The proof of inequality (7.4) and the remainder of the theorem now proceed just as in Theorem 4.7. □

Defining $\boldsymbol{\phi}_n^E$ and p_n^E as in § 5, we see that they again satisfy the partial differential equations (5.1) and (5.2). The boundary conditions now become

$$\boldsymbol{\phi}_n^E \cdot \boldsymbol{s} = 0, \qquad M_n\boldsymbol{\phi}_n^E = t^{n+1}D\frac{\partial \boldsymbol{\Phi}_{n-1}}{\partial n} \cdot \boldsymbol{n}, \qquad \frac{\partial p_n^E}{\partial n} = -t^{n-1}\frac{\partial p_{n-1}}{\partial n} - t^n\frac{\partial p_n}{\partial n},$$

and the variational equation (5.4) which enters the proof of Theorem 5.1 thus becomes

$$(C\,\mathcal{E}(\boldsymbol{\phi}_n^E), \mathcal{E}(\boldsymbol{\psi})) - (\mathbf{curl}\, p_n^E, \boldsymbol{\psi}) = -(\boldsymbol{R}_{n-2}, \boldsymbol{\psi}) + t^{n+1}D\left\langle \frac{\partial \boldsymbol{\Phi}_{n-1}}{\partial n} \cdot \boldsymbol{n}, \boldsymbol{\psi} \cdot \boldsymbol{n}\right\rangle,$$

valid for ψ with vanishing tangential component on $\partial\Omega$. We bounded the right-hand side of this equation in Theorem 7.2. This is the only additional consideration in establishing Theorem 5.1 in the hard simply-supported case.

The higher-order estimates in Theorem 5.2 also carry over to the present case, but again there is an additional term to be bounded because ϕ_n^E does not vanish on $\partial\Omega$. The bound for $\|\phi_n^E\|_s$ given in (5.5) must be modified to include the additional term

$$t^{n+1}\left|\frac{\partial \Phi_{n-1}}{\partial n}\cdot \boldsymbol{n}\right|_{s-3/2}.$$

In view of (7.1) and Theorem 4.3, this term is easily bounded by $t^n\|g\|_{s+n-3}$, which is no larger than other terms which were treated in the proof of Theorem 5.2. Of course, once Theorems 4.3, 5.1, and 5.2 are established, the regularity results given in Theorem 5.3 follow.

An asymptotic expansion and regularity results for the transverse displacement and the shear stress can be developed as in § 6. Naturally the boundary conditions in the defining problems for the expansion functions are changed. The boundary conditions on ϕ and p imply that $v = \omega - t^2 r$ satisfies, in addition to the differential equation $D\,\Delta^2 v = g$, the boundary conditions

$$v = 0, \qquad (1-\nu)\partial^2 v/\partial n^2 + \nu\,\Delta\, v = t^2(1-\nu)\kappa\partial p/\partial s \quad \text{on } \partial\Omega,$$

where κ denotes the curvature of $\partial\Omega$. It is then clear how to define the regular expansion for v and hence ω, and all the analysis of § 6 carries over easily.

Appendix. In this appendix we give the proof of Theorem 3.3 concerning the existence, uniqueness, and form of the solution of the boundary value problems defining the boundary correctors.

Proof. Differentiating (3.17) with respect to $\hat\rho$, we obtain

$$-\hat{\boldsymbol{A}}_5\cdot\frac{\partial^2\hat{\boldsymbol{\Phi}}_i}{\partial\hat\rho^2}+\frac{\partial^3\hat P_i}{\partial\hat\rho^3}=\frac{\partial\hat G_i(\hat\rho,\theta)}{\partial\hat\rho}.$$

Multiplying (3.16) by $\hat{\mathcal{A}}_0^{-1}$ and taking the inner product with $\hat{\boldsymbol{A}}_5$, we obtain

$$\hat{\boldsymbol{A}}_5\cdot\frac{\partial^2\hat{\boldsymbol{\Phi}}_i}{\partial\hat\rho^2}+\hat{\boldsymbol{A}}_5^t\hat{\mathcal{A}}_0^{-1}\hat{\boldsymbol{A}}_5\frac{\partial\hat P_i}{\partial\hat\rho}=-\hat{\boldsymbol{A}}_5^t\hat{\mathcal{A}}_0^{-1}\hat{\boldsymbol{F}}_i(\hat\rho,\theta).$$

Adding these equations and observing from Lemma 3.1 that $\hat{\boldsymbol{A}}_5^t\hat{\mathcal{A}}_0^{-1}\hat{\boldsymbol{A}}_5=-c^2$, we get

$$\frac{\partial^3\hat P_i}{\partial\hat\rho^3}-c^2\frac{\partial\hat P_i}{\partial\hat\rho}=-\hat{\boldsymbol{A}}_5^t\hat{\mathcal{A}}_0^{-1}\hat{\boldsymbol{F}}_i(\hat\rho,\theta)+\frac{\partial\hat G_i(\hat\rho,\theta)}{\partial\hat\rho}=:\hat H_i. \tag{8.1}$$

The general solution of the associated homogeneous equation is $c_1(\theta)+c_2(\theta)e^{-c\hat\rho}+c_3(\theta)e^{c\hat\rho}$, with the functions c_i arbitrary. Now if we have two solutions to (3.16)–(3.19), then the difference in the values of $\hat P_i$ must be of this form. Applying (3.19) implies that c_1 and c_3 vanish, and then the homogeneous form of (3.18) implies that c_2 vanishes. Thus there can be at most one function $\hat P_i$ satisfying (3.16)–(3.19). Once $\hat P_i$ is known, $\hat{\boldsymbol{\Phi}}_i$ is determined up to the addition of a function linear in $\hat\rho$ by (3.16).

In light of (3.19), $\hat{\boldsymbol{\Phi}}_i$ is uniquely determined. Thus we have shown that there can be at most one solution $(\hat{\boldsymbol{\Phi}}_i, \hat{P}_i)$ to (3.16)–(3.19).

Let us say that a scalar-valued function $\hat{Q}(\hat{\rho}, \theta)$ is of type (m, i) if

$$\hat{Q}(\hat{\rho}, \theta) = e^{-c\hat{\rho}} \sum_{k=0}^{m} \sum_{j=0}^{i} \sum_{l=0}^{i-j} \alpha_{jkl}(\theta) \hat{\rho}^k \frac{\partial^l}{\partial \theta^l} \widetilde{\frac{\partial p_j}{\partial n}}(0, \theta)$$

for some smooth functions $\alpha_{jkl}(\theta)$. A vector-valued function is of type (m, i) if all components are. We claim that there is a solution $(\hat{\boldsymbol{\Phi}}_i, \hat{P}_i)$ to (3.16)–(3.19) which is of of type (i, i). We will establish the claim by induction on i, thereby completing the proof of the theorem. The solution given in (3.20) verifies the claim for $i = 0$. Now suppose that $(\hat{\boldsymbol{\Phi}}_j, \hat{P}_j)$ is of type (j, j) for $j = 0, 1, \cdots, i-1$. It follows easily from their respective definitions (just after (3.16) and in (3.17) and (8.1)) that $\hat{\boldsymbol{F}}_i$, $\hat{G}_i$, and $\hat{H}_i$ are of type $(i-1, i)$. It is then elementary to see that the differential equation (8.1) has a unique solution of type (i, i) satisfying the boundary condition (3.18). Next, there is a unique function $\hat{\boldsymbol{\Phi}}_i$ of type (i, i) satisfying (3.16). Together (3.16), (8.1), and the decay at infinity of $\hat{\boldsymbol{\Phi}}_i$, $\hat{G}_i$, and the $\hat{\rho}$-derivatives of $\hat{P}_i$ imply (3.17). Thus $(\hat{\boldsymbol{\Phi}}_i, \hat{P}_i)$ satisfy (3.16)–(3.19) and are of the desired form. This completes the induction.

REFERENCES

[1] D. N. ARNOLD AND R. S. FALK, *A uniformly accurate finite element method for the Reissner–Mindlin plate*, SIAM J. Numer. Anal., 26 (1989), pp. 1276–1290.

[2] T. C. ASSIFF AND D. H. Y. YEN, *On the solutions of clamped Reissner-Mindlin plates under transverse loads*, Quart. Appl. Math., 45 (1987), pp. 679–690.

[3] F. BREZZI AND M. FORTIN, *Numerical approximation of Mindlin-Reissner plates*, Math. Comp., 47 (1986), pp. 151–158.

[4] F. BREZZI, *On the existence, uniqueness, and approximation of saddle point problems arising from Lagrangian multipliers*, RAIRO Anal. Numér., 2 (1974), pp. 129–151.

[5] L. CATTABRIGA, *Su un problema al contorno relativo al sistema di equazioni di Stokes*, Rend. Sem. Mat. Padova, 31 (1961), pp. 1–33.

[6] B. M. FRAEIJS DE VEUBEKE, *A Course in Elasticity*, Springer-Verlag, Berlin, New York, 1979.

[7] B. HÄGGBLAD AND K-J. BATHE, *Specifications of boundary conditions for Mindlin/Reissner theory based on plate bending finite elements*, preprint.

[8] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications, I*, Springer-Verlag, Berlin, New York, 1972.

[9] R. D. MINDLIN, *Influence of rotary inertia and shear on flexural motions of elastic plates*, Trans. ASME Ser. E J. Appl. Mech., 18 (1958), pp. 31–38.

[10] E. REISSNER, *The effect of transverse shear deformation on the bending of elastic plates*, Trans. ASME Ser. E J. Appl. Mech., 67 (1945), pp. A69–A77.

[11] F. Y. M. WAN, *Lecture Notes on Problems in Elasticity:* II. *Linear Plate Theory*, Institute of Applied Mathematics, University of British Columbia, Tech. Rep. 83-15, Sept., 1983.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 313-326, March 1990
© 1990 Society for Industrial and Applied Mathematics
002

A SINGULAR PERTURBATION ANALYSIS OF REVERSE BIASED *pn*-JUNCTIONS*

CHRISTIAN SCHMEISER†

Abstract. A two-dimensional version of the drift-diffusion model for stationary flow of charge carriers in semiconductor devices is considered. It consists of a system of three elliptic equations that is singularly perturbed for large applied biases. For the case of a *pn*-junction diode under strong reverse bias, an approximating problem, which includes a free-boundary problem for the potential and a mixed elliptic-hyperbolic problem for the analysis of current flow, is derived. The solvability of this formal limit problem is proved.

Key words. semiconductors, drift-diffusion model, singular perturbations

AMS(MOS) subject classifications. 35B25, 35R35

1. Introduction.

In this paper we consider the system

$$\Delta\psi = n - p - C(x), \tag{1.1a}$$

$$\varepsilon\nabla n - n\nabla\psi = \varepsilon J_n, \qquad \operatorname{div} J_n = 0, \tag{1.1b}$$

$$-\varepsilon\nabla p - p\nabla\psi = \varepsilon J_p, \qquad \operatorname{div} J_p = 0 \tag{1.1c}$$

for $x \in \Omega \subset \mathbb{R}^2$, where Ω is a bounded simply-connected domain with Lipschitz boundary representing the semiconductor part of an electronic device. The scaling that leads to (1.1) (see [13]) is valid for large applied biases. The dimensionless parameter ε is small and positive in this case. The Poisson equation (1.1a) determines the electrostatic potential ψ and thus the electric field $E = -\nabla\psi$. The term $-n+p+C(x)$ on the right-hand side is the space charge density with contributions from the negatively charged free electrons (density n), the positively charged holes (density p), and a fixed distribution of charges $C(x)$ called the doping profile. The equations in (1.1b, c) represent a convection-diffusion model for the electron and hole current densities J_n and J_p, and current continuity.

We consider (1.1) subject to mixed Dirichlet-Neumann boundary conditions. Dirichlet conditions are given at contacts (disjoint, connected boundary segments, closed with respect to $\partial\Omega$) and homogeneous Neumann conditions at the remaining insulating part of the boundary.

A semiconductor device is specified by the doping profile and the number and location of contacts. We consider a *pn*-junction diode, where Ω splits into a *p*-region Ω_1 where $C(x) < 0$ holds and an *n*-region Ω_2 where $C(x) > 0$ holds. Ω_1 and Ω_2 are separated by the *pn*-junction Γ. The following technical assumptions will be used:

(A1) Γ is a smooth curve that meets $\partial\Omega$ under right angles. In neighborhoods of the points where Γ meets $\partial\Omega$, the boundary is given by straight line segments.

(A2) The doping profile is smooth in $\Omega_1 \cup \Omega_2$ and has jump discontinuities at Γ. $|C(x)| \geqq \gamma > 0$ holds.

* Received by the editors August 3, 1988; accepted for publication (in revised form) May 26, 1989. This work was supported by "Österreichischer Fonds zur Förderung der wissenschaftlichen Forschung."

† Institut für Angewandte und Numerische Mathematik, Technical University of Vienna, Wiedner Haupstrasse 8-10/115, A-1040 Vienna, Austria.

A diode has two contacts C_1, $C_2 \subset \partial\Omega$ with $C_1 \subset \partial\Omega_1$ and $C_2 \subset \partial\Omega_2$. We assume the following:

(A3) The closure $\bar{\Gamma}$ of the *pn*-junction does not intersect the contacts C_1 and C_2.

A typical example for the device geometry is depicted in Fig. 1.1.

We consider (1.1) subject to the boundary conditions

$$\psi|_{C_1} = 0, \qquad \psi|_{C_2} = \alpha,$$

$$n|_{C_1,C_2} = \tfrac{1}{2}(C + \sqrt{C^2 + 4\delta^4})|_{C_1,C_2}, \tag{1.2}$$

$$p|_{C_1,C_2} = \tfrac{1}{2}(-C + \sqrt{C^2 + 4\delta^4})|_{C_1,C_2},$$

$$\nabla\psi \cdot \nu|_{\partial\Omega_N} = \nabla n \cdot \nu|_{\partial\Omega_N} = \nabla p \cdot \nu|_{\partial\Omega_N} = 0 \tag{1.3}$$

where $\partial\Omega_N = \partial\Omega \backslash (C_1 \cup C_2)$ denotes the insulating boundary segments and ν the unit outward normal.

The parameter α represents the applied voltage with $\alpha > 0$ in the case of large reverse bias. The Dirichlet data for n and p are obtained from the assumptions of vanishing space charge ($n - p - C = 0$) and thermal equilibrium ($np = \delta^4$) at the contacts. The thermal equilibrium equation represents a mass-action law, where δ^2 is the scaled intrinsic number of the semiconductor (see [15] for details). In practical applications δ^2 takes small values because densities (such as the intrinsic number) are scaled by the maximal value of the doping profile, which usually is much larger than the intrinsic carrier density.

Several existence proofs for (1.1)-(1.3) can be found in the literature (see [11] and [12] and references therein). An application of Theorem 3.2.1 in [11] yields Theorem 1.1.

THEOREM 1.1. *Problem* (1.1)-(1.3) *has a solution* $(\psi, n, p) \in (H^1(\Omega) \cap L^\infty(\Omega))^3$ *that satisfies*

$$-\varepsilon\beta \leqq \psi \leqq \alpha + \varepsilon\beta \quad \text{in } \Omega$$

where

$$\beta := 2 \ln [(\|C\|_{L^\infty(\Omega)} + \sqrt{\|C\|^2_{L^\infty(\Omega)} + 4\delta^4})/2\delta^2]$$

holds.

Although β tends to infinity as $\delta \to 0$, the product $\varepsilon\beta$ is usually small compared to unity. This suggests an asymptotic analysis of (1.1)-(1.3) for $\varepsilon \to 0$. Theorem 1.1

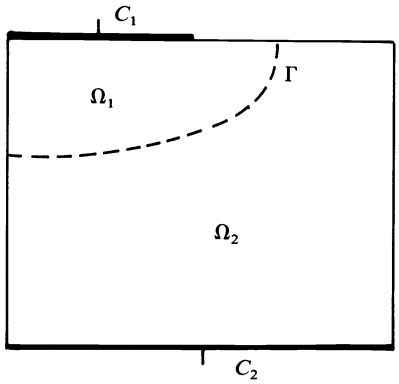


FIG. 1.1. *Device geometry.*

shows that the potential is bounded uniformly in ε. For a rigorous analysis of the limit $\varepsilon \to 0$ additional a priori estimates of this kind would be necessary. However, the existence result on which Theorem 1.1 is based does not provide sufficiently sharp bounds.

In [4] Caffarelli and Friedman treated related problems. The main difference from (1.1)-(1.3) is a modification of the boundary conditions that allows for a rigorous analysis of a simplified problem.

Similar results for one-dimensional problems have been developed in a series of papers by Brezzi et al. [1]-[3]. The analysis of the present work is based on the formal methods of singular perturbation theory.

Although the modified boundary conditions in the above-mentioned papers do not allow for an analysis of current flow, the limiting behavior of the potential is obtained. Under appropriate assumptions ψ is shown to converge to the solution of the variational inequality

$$\psi_0 \in K_\alpha : \int_\Omega \nabla\psi_0 \nabla(v-\psi_0)\,dx \geqq \int_\Omega C(x)(v-\psi_0)\,dx \quad \forall v \in K_\alpha \tag{1.4}$$

where

$$K_\alpha := \{\psi \in H^1(\Omega) : \psi|_{C_1} = 0,\ \psi|_{C_2} = \alpha,\ 0 \leqq \psi \leqq \alpha \text{ a.e. in } \Omega\} \tag{1.5}$$

holds. Inequality (1.4) can be interpreted as a double obstacle problem for the deformation ψ of a membrane. The membrane lies between the obstacles represented by $\psi = 0$ and $\psi = \alpha$ and is fixed to the lower obstacle along C_1 and to the upper obstacle along C_2. In this context $C(x)$ is a transversal force pushing downward in Ω_1 and upward in Ω_2.

Without being derived as a limit of (1.1)-(1.3), the double obstacle problem (1.4) has already been formulated as a model for the potential distribution in [7] and [10]. It can be motivated by the reduced equations

$$\Delta\psi_0 = n_0 - p_0 - C(x), \qquad n_0 \nabla\psi_0 = p_0 \nabla\psi_0 = 0 \tag{1.6}$$

and the estimates in Theorem 1.1.

For the carrier densities n_0 and p_0 the reduced system implies

$$n_0 - p_0 - C(x) = 0 \quad \text{in } Z \cup A$$

where Z and A denote the coincidence sets

$$Z := \{x \in \Omega : \psi_0(x) = 0\},$$

$$A := \{x \in \Omega : \psi_0(x) = \alpha\},$$

$$n_0 = p_0 = 0 \quad \text{in } N = \Omega \setminus (Z \cup A).$$

In the physical literature the noncoincidence set N is called the depletion region, as it is depleted of charge carriers, or the space charge region.

The equations above are not sufficient for characterizing the limiting charge carrier and current densities. For a one-dimensional situation the limiting problem has been completed by Schmeiser [13]. It is the main purpose of the present work to extend these results to the two-dimensional case.

The formulation of the limit problem requires certain topological properties of the sets Z, A, and N that are proved in § 2 to hold for α small enough.

In § 3 we introduce asymptotic expansions in powers of ε for the solution that allow us to formulate equations determining the current flow. It turns out that in the

space charge region N the flow is purely convective and governed by a hyperbolic system, whereas diffusion is significant in Z and A, where a system of elliptic equations has to be solved. A proof of existence of a locally unique solution of the coupled problem for small δ^4 is the main result of § 3.

2. The double obstacle problem. This section is concerned with an analysis of problem (1.4) for small values of α. Standard results for variational inequalities [8] yield Theorem 2.1.

THEOREM 2.1. (A) *Problem* (1.4) *has a unique solution* ψ_0 *that is Hölder continuous. Besides,* $\psi_0 \in W^{2,p}(\Omega')$ *holds for any* $p < \infty$ *and for any subdomain* Ω' *of* Ω *whose closure does not contain critical boundary points, i.e.,* (a) *points where* $\partial\Omega$ *is not smooth, and* (b) *edges of the contacts where Dirichlet and Neumann boundary conditions meet.*

$$\text{(B)} \qquad \Delta\psi_0 + C(x) \leqq 0 \quad \text{a.e. in } Z \cup N,$$

$$\Delta\psi_0 + C(x) \geqq 0 \quad \text{a.e. in } A \cup N.$$

A direct consequence of (B) is Corollary 2.1.

COROLLARY 2.1. $Z \subset \bar{\Omega}_1$, $A \subset \bar{\Omega}_2$, $\Gamma \subset \bar{N}$. By continuity of ψ_0, Z and A are closed, and thus, N is open relative to Ω. A refinement of the last result in Corollary 2.1 is Lemma 2.1.

LEMMA 2.1. $\Gamma \subset N$.

Proof. Suppose $x_0 \in \Gamma \cap A$. The smoothness of Γ and the continuity of ψ_0 imply the existence of a ball $B_R(y) \subset \Omega_1 \cap N$ with $x_0 \in \partial B_R(y)$ (see Fig. 2.1). We have

(a) $\Delta\psi_0 = -C(x) > 0$ in $B_R(y)$,

(b) $\alpha = \psi_0(x_0) > \psi_0(x)$ for $x \in B_R(y)$,

(c) ψ_0 continuous.

Application of Lemma 3.4 in [6] leads to

$$\nabla\psi_0(x_0) \cdot \nu > 0$$

where ν is the outward unit normal on $\partial B_R(y)$. This is in contradiction to $\nabla\psi_0 = 0$ in A. Analogously we prove $\Gamma \cap Z = \{\ \}$, which completes the proof.

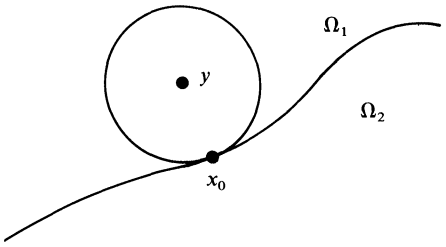


FIG. 2.1

The lemma immediately implies $\Gamma_Z \subset \Omega_1$, $\Gamma_A \subset \Omega_2$ for the free boundaries Γ_Z separating Z and N and Γ_A separating A and N. From the smoothness of $C(x)$ in $\Omega_1 \cup \Omega_2$ we obtain smoothness of Γ_Z and Γ_A by standard results (see [5], [8]).

For the analysis of the following section we will need Z, A, and N to be connected. For Z and A this cannot be expected to hold in general. But we have Lemma 2.2.

LEMMA 2.2. *N is connected.*

Proof. By the above results there is a component N_Γ of N with $\Gamma \subset N_\Gamma$. Assume the existence of an additional component $\tilde{N} \subset \Omega_1$. We have

$$\Delta\psi_0 = -C(x) > 0 \quad \text{in } \tilde{N},$$
$$\psi_0 = 0 \quad \text{or} \quad \nabla\psi_0 \cdot \nu = 0 \quad \text{on } \partial\tilde{N},$$

which implies $\psi_0 \leqq 0$ in $\tilde{N}$ by the maximum principle. This is in contradiction to $\psi_0 > 0$ in N. Analogously we show that there does not exist an additional component of N in Ω_2. Thus, $N = N_\Gamma$, which completes the proof.

In [13] we have shown that Γ_Z and Γ_A converge to Γ as $\alpha \to 0$ in the one-dimensional case. In the following this result is extended to two dimensions.

We define

$$\Omega_\rho := \{x \in \Omega : d(x, \Gamma) < \rho\}.$$

With $s(x) := d(x, \Gamma)$ we have

$$s \in C^2(\Omega_{\rho_0}) \quad \text{for some } \rho_0 > 0,$$
$$|\nabla s| = 1, \quad \Delta s \geqq -M \quad \text{in } \Omega_{\rho_0} \quad \text{for some } M > 0$$

by assumption (A1).

THEOREM 2.2. *Let α be small enough so that $\rho = (2\alpha(M\rho_0 + 1)/\gamma)^{1/2} \leq \rho_0$ holds. Then $N \subset \Omega_\rho$ holds.*

Remark. The theorem is a mathematical version of the statement, "The width of the depletion region is proportional to the square root of the applied bias," which is well known in the engineering literature (see [16]) for the case of abrupt junctions. The result could be extended to graded junctions where the doping profile smoothly changes sign across Γ and, accordingly, $\gamma > 0$ with the required properties does not exist. In this case the width of the depletion region is $O(\alpha^{1/3})$ for small values of α.

Proof. Let us deal with the part of N lying in Ω_1 first. We define a comparison function by

$$w(x) := \begin{cases} 0, & x \in \Omega_1 \setminus \Omega_\rho, \\ \alpha(1 - s(x)/\rho)^2, & x \in \Omega_1 \cap \Omega_\rho. \end{cases}$$

Obviously, $w \in C^1(\Omega_1)$, $w \geqq 0$ in Ω_1, $w|_{C_1} = 0$, $w|_\Gamma = \alpha$ holds. Besides, we have for $x \in \Omega_1 \cap \Omega_\rho$

$$\Delta w(x) = 2\alpha/\rho^2[(s - \rho)\Delta s + |\nabla s|^2]$$
$$\leqq \frac{\gamma}{M\rho_0 + 1}[(\rho - s)M + 1]$$
$$\leqq \gamma \frac{\rho M + 1}{\rho_0 M + 1} \leqq \gamma \leqq -C(x),$$

and for $x \in \Omega_1 \setminus \Omega_\rho$

$$\Delta w(x) = 0 < -C(x).$$

This shows that w is a supersolution of the equation $\Delta w + C(x) = 0$ in Ω_1. A comparison principle as in [8] implies

$$\psi_0 \leqq w \quad \text{in } \Omega_1,$$

and thus, $\psi_0 = 0$ in $\Omega_1 \setminus \Omega_\rho$. Similarly, it can be shown that $\psi_0 = \alpha$ in $\Omega_2 \setminus \Omega_\rho$ holds, which completes the proof.

COROLLARY 2.2. *$\psi_0 \in W^{2,p}(\Omega)$ for $p < \infty$ and for α small enough.*

Proof. $\psi_0 = 0$ or $\psi_0 = \alpha$ in neighborhoods of critical boundary points by Theorem 2.2 and assumption (A1). The result now follows by Theorem 2.1.

Theorem 2.2 shows that the free boundaries meet $\partial\Omega$ within the Neumann segments for α small enough.

LEMMA 2.3. (a) *If $\partial\Omega$ is smooth in a neighborhood of a point x_0 where one of the free boundaries meets $\partial\Omega_N$, then the angle between the free boundary and $\partial\Omega_N$ at x_0 is $\pi/2$.*

(b) $\nabla\psi_0/|\nabla\psi_0|$ *is orthogonal to the free boundaries.*

Proof. (a) A locally orthogonal smooth coordinate transformation moves x_0 to the origin and a neighborhood of x_0 in Ω to $\{x_2>0\}\cap B_R(0)$, where (x_1, x_2) are the new coordinates. (See Fig. 2.2.). The extension of ψ_0 to $B_R(0)$ as an even function of x_2 is the solution of a variational inequality in $B_R(0)$. The smoothness of the free boundary that follows from standard results, and its symmetry with respect to the x_1-axis, imply the conclusion of the lemma.

(b) Let $x_0\in\Omega$ lie on Γ_Z. Then ψ_0 is smooth in a neighborhood of x_0 in N by elliptic regularity. As in the proof of (a) we transform this neighborhood to $\{x_2>0\}\cap B_R(0)$. We have $\psi_0(x_1, x_2)=x_2{}^2 f(x_1, x_2)$ with f smooth because $\psi_0=0$ for $x_2<0$ and $\psi_0\in C^1(B_R)$. Assumption (A2) and the fact that ψ_0 satisfies $\Delta\psi_0+C(x)=0$ in N imply

$$f(x_1, 0)\geqq\tilde{\gamma}>0.$$

If we use this, straightforward computation shows

$$\lim_{x_2\to 0}\nabla\psi_0/|\nabla\psi_0|=(0, 1),$$

which is the desired result. It can be extended to points where Γ_Z meets the boundary by reflection as in the proof of Lemma 2.3. Points on Γ_A are treated analogously.

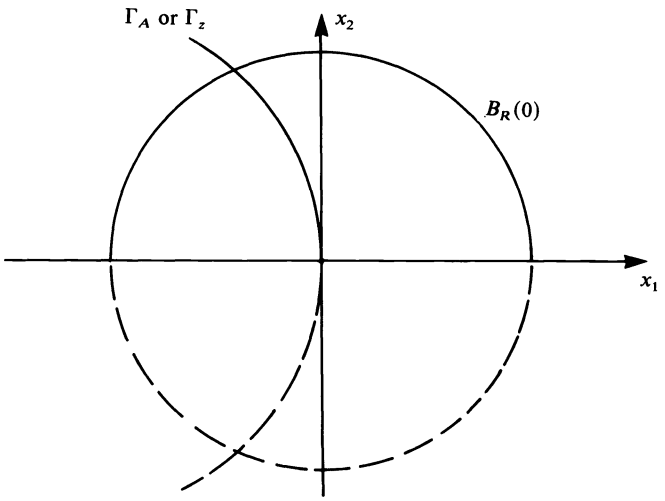


FIG. 2.2

The remaining part of this section is devoted to a more detailed analysis of the solution of (1.4) for small α. With the substitution $\psi_0=\alpha\phi$, (1.4) changes to

$$\phi\in K_1:\ \alpha\int_\Omega\nabla\phi\nabla(v-\phi)\,dx\geqq\int_\Omega C(x)(v-\phi)\,dx\quad\forall v\in K_1 \tag{2.1}$$

where K_1 is defined as in (1.5). This is a singularly perturbed variational inequality. The reduced problem obtained by formally setting $\alpha = 0$ in (2.1) does not have a solution. This difficulty can be overcome by considering a larger solution space. Lions [9] treated problems where the reduced variational inequality is governed by a bilinear form that is coercive in the enlarged space. As this is not possible in our situation, we guarantee solvability by choosing the larger space such that K_1 becomes bounded. Our choice is $L^2(\Omega)$ and the reduced problem is defined as

$$\bar{\phi} \in \bar{K}_1 : 0 \geqq \int_\Omega C(x)(v - \bar{\phi})\, dx \quad \forall v \in \bar{K}_1 \tag{2.2}$$

where $\bar{K}_1$ is the closure of K_1 in $L^2(\Omega)$. Obviously (2.2) has the unique solution

$$\bar{\phi} = \begin{cases} 0 & \text{a.e. in } \Omega_1, \\ 1 & \text{a.e. in } \Omega_2. \end{cases} \tag{2.3}$$

From the proof of Theorem 2.2 we obtain a validity result for the formal approximation $\bar{\phi}$.

LEMMA 2.4. *For the solutions ϕ of* (2.1) *and $\bar{\phi}$ of* (2.2)

$$\|\phi - \bar{\phi}\|_{L^p(\Omega)} = O(\alpha^{1/2p}) \quad \text{as } \alpha \to 0$$

holds for $p \geqq 1$.

Proof. For the function $w(x)$ defined in the proof of Theorem 2.2 we have

$$|\phi(x) - \bar{\phi}(x)| \leqq w/\alpha \quad \text{a.e. in } \Omega_1.$$

This implies

$$\begin{aligned} \|\phi - \bar{\phi}\|_{L^p(\Omega_1)} &\leqq \left(\int_{\Omega_1 \cap \Omega_\rho} (1 - s(x)/\rho)^{2p}\, dx \right)^{1/p} \\ &\leqq (\text{meas } (\Omega_1 \cap \Omega_\rho))^{1/p} = O(\alpha^{1/2p}). \end{aligned}$$

Similarly we obtain

$$\|\phi - \bar{\phi}\|_{L^p(\Omega_2)} = O(\alpha^{1/2p}),$$

which completes the proof.

Obviously, $\bar{\phi}$ is not smooth enough for a uniformly valid approximation. To improve on that we introduce a correction layer at Γ. Considering a local change of variables

$$(x_1, x_2) \to (r, s)$$

where

$$s(x) = \begin{cases} -d(x, \Gamma) & \text{in } \Omega_1, \\ d(x, \Gamma) & \text{in } \Omega_2, \end{cases}$$

we introduce the fast variable $\sigma = s/\sqrt{\alpha}$ and the σ-interval $I = (-\sqrt{2(M\rho_0 + 1)/\gamma}, \sqrt{2(M\rho_0 + 1)/\gamma})$ (compare with Theorem 2.2). Then the layer problem is given by the one-dimensional variational inequality:

$$\hat{\phi} \in \tilde{K} : \int_I \hat{\phi}_\sigma (v_\sigma - \hat{\phi}_\sigma)\, d\sigma \geqq \int_I C(r, \sigma)(v - \hat{\phi})\, d\sigma \quad \forall v \in \tilde{K}$$

where $\tilde{K}$ is defined by

$$\begin{aligned} \tilde{K} = \{ \hat{\phi} \in H^1(I) : \hat{\phi}(-\sqrt{2(M\rho_0 + 1)/\gamma}) = 0, \\ \hat{\phi}(\sqrt{2(M\rho_0 + 1)/\gamma}) = 1, 0 \leqq \hat{\phi} \leqq 1 \}. \end{aligned}$$

The one-dimensional variational inequality has been treated in [13]. It is easy to show that the free boundaries are located in the interior of I and that $\hat{\phi}$ is strictly monotonically increasing between them.

A formal approximation of the solution of (2.1) is defined by

$$\phi_{as} = \begin{cases} \bar{\phi} & \text{in } \Omega \backslash \Omega_\rho, \\ \hat{\phi} & \text{in } \Omega_\rho. \end{cases}$$

THEOREM 2.3. $\|\phi - \phi_{as}\|_{H^1(\Omega)} = O(\alpha^{1/4})$.

Proof. By assumptions (A1), (A2) it is straightforward to show that ϕ_{as} solves the variational inequality

$$(2.4) \quad \phi_{as} \in K_1: \alpha \int_\Omega \nabla \phi_{as}(v - \phi_{as})\, dx \geqq \int_\Omega (C(x) + \alpha g(x))(v - \phi_{as})\, dx \quad \forall v \in K_1$$

where $\|g\|_{L^\infty(\Omega)}$ is bounded independently of α. For $w = \phi - \phi_{as}$ we get from Lemma 2.4 and the definition of ϕ_{as}

$$(2.5) \qquad \|w\|_{L^p(\Omega)} = O(\alpha^{1/2p}).$$

If we set $v = \phi_{as}$ in (2.1) and $v = \phi$ in (2.4), the sum of the resulting inequalities reads

$$-\alpha \int_\Omega |\nabla w|^2\, dx \geqq \alpha \int_\Omega g(x) w\, dx.$$

This implies

$$\|\nabla w\|^2_{L^2(\Omega)} \leqq \|g\|_{L^\infty(\Omega)} \|w\|_{L^1(\Omega)} = O(\alpha^{1/2})$$

by Hölder's inequality, the boundedness of $g(x)$, and (2.5) with $p = 1$. Combining this estimate with (2.5) ($p = 2$) completes the proof.

Finally, we state two assumptions that are needed in the following section:

(A4) Z and A are connected.

(A5) $\nabla \psi_0 / |\nabla \psi_0|$ is Lipschitz in N.

Since both assumptions are satisfied by the formal approximations constructed above, the author conjectures that they are satisfied for α small enough, although a proof of this conjecture is not available.

3. Analysis of current flow. By formally setting $\varepsilon = 0$ in (1.1) and considering the solution of (1.4) we obtain the equations

$$(3.1a) \qquad n_0 - p_0 - C(x) = 0 \qquad \text{in } Z \cup A,$$

$$(3.1b) \qquad n_0 = p_0 = 0 \qquad \text{in } N,$$

$$(3.1c) \qquad \operatorname{div} J_{n0} = \operatorname{div} J_{p0} = 0 \qquad \text{in } \Omega,$$

which are not sufficient for the computation of n_0, p_0, J_{n0}, and J_{p0}. It is a standard procedure in perturbation theory to derive additional equations by introducing asymptotic expansions for the solution and equating coefficients of higher-order terms (see [14], [18]). We make the ansatz

$$w = w_0 + \varepsilon w_1 + O(\varepsilon^2)$$

where w stands for any of the solution components. Substitution in (1.1b, c) and comparison of coefficients of ε leads to

$$\nabla n_0 - n_0 \nabla \psi_1 - n_1 \nabla \psi_0 = J_{n0},$$
$$-\nabla p_0 - p_0 \nabla \psi_1 - p_1 \nabla \psi_0 = J_{p0}.$$

With (3.1) and our knowledge about ψ_0 we arrive at

$$J_{n0} = -n_1 \nabla \psi_0, \qquad J_{p0} = -p_1 \nabla \psi_0 \quad \text{in } N, \tag{3.2}$$

$$\begin{aligned} J_{n0} &= \nabla n_0 - n_0 \nabla \psi_1 \\ J_{p0} &= -\nabla p_0 - p_0 \nabla \psi_1 \end{aligned} \quad \text{in } Z \cup A. \tag{3.3}$$

Our next aim is to eliminate ψ_1 from (3.1a), (3.3). In [13] this is done by introducing the slow variable np. As the resulting problem is difficult to handle in the two-dimensional case, we instead introduce the change of variables

$$n_0 = e^{\psi_1} u, \qquad p_0 = e^{-\psi_1} v.$$

If we use (3.1a) for eliminating ψ_1, (3.3) changes to

$$\begin{aligned} J_{n0} &= \frac{C + \sqrt{C^2 + 4uv}}{2u} \nabla u \\ J_{p0} &= -\frac{-C + \sqrt{C^2 + 4uv}}{2v} \nabla v \end{aligned} \quad \text{in } Z \cup A. \tag{3.4}$$

Substituting (3.2) and (3.4) in (3.1c) yields the system

$$\operatorname{div}(n_1 \nabla \psi_0) = \operatorname{div}(p_1 \nabla \psi_0) = 0 \quad \text{in } N, \tag{3.5a}$$

$$\operatorname{div}\left(\frac{C + \sqrt{C^2 + 4uv}}{2u} \nabla u\right) = \operatorname{div}\left(\frac{-C + \sqrt{C^2 + 4uv}}{2v} \nabla v\right) = 0 \quad \text{in } Z \cup A, \tag{3.5b}$$

which is elliptic in $Z \cup A$ and hyperbolic in N.

Note that the computation of the current densities J_{n0}, J_{p0} in N necessitates the determination of the $O(\varepsilon)$ corrections n_1 and p_1 of the charge carrier densities. For singular singularly perturbed problems, it is typical to have to deal with terms of different orders simultaneously.

The formulation of the problem is completed by prescribing boundary conditions at $\partial\Omega$ and interface conditions at Γ_Z and Γ_A (see Fig. 3.1). As ψ_0 satisfies the original

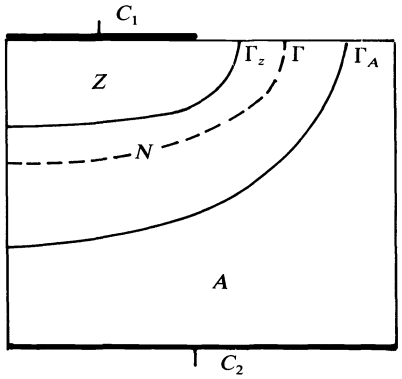


FIG. 3.1

boundary conditions for ψ, homogeneous boundary conditions for the correction ψ_1 are required. This leads to

$$u|_{C_1,C_2} = \tfrac{1}{2}(C + \sqrt{C^2+4\delta^4})|_{C_1,C_2},$$
$$v|_{C_1,C_2} = \tfrac{1}{2}(-C + \sqrt{C^2+4\delta^4})|_{C_1,C_2}, \tag{3.6a}$$

$$\nabla u \cdot \nu, \nabla v \cdot \nu|_{(\partial Z \cup \partial A)\cap \partial\Omega_N} = 0. \tag{3.6b}$$

Since the characteristics of (3.5a) are parallel to $\partial N \cap \partial\Omega_N$, we do not need any boundary conditions there. At the interfaces Γ_Z and Γ_A, we obviously require continuity of the normal components of the current densities

$$[J_{n0} \cdot \nu]_{\Gamma_Z,\Gamma_A} = [J_{p0} \cdot \nu]_{\Gamma_Z,\Gamma_A} = 0 \tag{3.6c}$$

where $[\,\cdot\,]_S$ denotes the jump of a quantity across the curve S and ν stands for the unit normal to $\Gamma_Z(\Gamma_A)$ pointing outward from $Z(A)$. The above-mentioned observation that np is a slow variable leads to the condition $[n_0 p_0]_{\Gamma_Z,\Gamma_A} = 0$.

By requiring the concentrations to be nonnegative, we see from (3.1) that only p_0 has a jump discontinuity at Γ_Z and only n_0 has a jump discontinuity at Γ_A. This implies that for u and v

$$u|_{\Gamma_Z} = 0, \qquad v|_{\Gamma_A} = 0. \tag{3.6d}$$

The reduced solution is determined completely by solving (3.5), (3.6).

Before starting to analyze the reduced problem we mention that the jumps of p_0 at Γ_Z and n_0 at Γ_A can be smoothed by layer solutions. They are obtained by introducing fast local coordinates along Γ_Z and Γ_A. This leads to a system of ordinary differential equations on $\mathbb{R}$ that has been analyzed in [13].

As a preliminary step in the discussion of the reduced problem we consider the hyperbolic equations (3.5a). As $\nabla\psi_0$ vanishes at Γ_Z and Γ_A, the solutions n_1 and p_1 have singularities there in the case of nonvanishing current densities. It has been demonstrated in [13] that, despite these singularities, matching to layer solutions can be carried out. The following lemma shows that

$$n_1, p_1 = O(|\nabla\psi_0|^{-1})$$

holds close to Γ_Z and Γ_A. From the proof of Lemma 2.3 we see that $|\nabla\psi_0|$ is a possible choice of a local variable there. Written in the fast variable $\xi = |\nabla\psi_0|/\sqrt{\varepsilon}$, $O(\sqrt{\varepsilon}/\xi)$-contributions result from n_1 and p_1; these can be matched to algebraically decaying layer solutions of order $O(\sqrt{\varepsilon})$. The current flow through N is analyzed in Lemma 3.1.

LEMMA 3.1. *Let the operator* $S: H^{1/2}(\Gamma_Z) \to H^{1/2}(\Gamma_A)$ *be defined by*

$$S(f) := n\nabla\psi_0 \cdot \nu|_{\Gamma_A}$$

where n is the solution of

$$\operatorname{div}(n\nabla\psi_0) = 0 \quad \text{in } N,$$
$$n\nabla\psi_0 \cdot \nu|_{\Gamma_Z} = f.$$

Then S is a bounded and boundedly invertible linear operator.

Proof. We introduce $\hat{n} = n|\nabla\psi_0|$, which must solve the Cauchy problem

$$\operatorname{div}(\hat{n}\nabla\psi_0/|\nabla\psi_0|) = 0 \quad \text{in } N,$$
$$\hat{n}|_{\Gamma_Z} = f$$

because $\nabla\psi_0 \cdot \nu/|\nabla\psi_0||_{\Gamma_Z} = 1$ by Lemma 2.3(b). Assumption (A5) implies existence and uniqueness of $\hat{n}$.

We define the mapping $\phi : \Gamma_Z \to \Gamma_A$ by following characteristics starting at $x \in \Gamma_Z$ until they leave N at $\phi(x) \in \Gamma_A$. It is another consequence of assumption (A5) that ϕ is a diffeomorphism.

Let $x(s)$ denote a parametrization of Γ_Z by arclength. An application of the divergence theorem to the region bounded by the segment between $x(s)$ and $x(s+h)$ on Γ_Z, the characteristics starting at $x(s)$ and $x(s+h)$, and the segment between $\phi(x(s))$ and $\phi(x(s+h))$ on Γ_A yields

$$\int_s^{s+h} f(x(\sigma))\, d\sigma = -\int_s^{s+h} S(f)(\phi(x(\sigma)))|\phi_x \dot{x}|\, d\sigma.$$

By letting $h \to 0$ we obtain the representation

$$f(x) = -S(f)(\phi(x))|\phi_x \dot{x}|$$

for the inverse of S. A final application of assumption (A5) and Lemma 2.3(b) implies $|\phi_x \dot{x}| > 0$, which completes the proof.

Remark. The space $H^{1/2}$ was chosen only for convenience in the proof of Theorem 3.1 below. Obviously the result also holds for other function spaces.

Using Lemma 3.1 and introducing $u|_A = e^w$, $v|_Z = e^z$, we must still solve a system of four coupled elliptic boundary value problems:

$$\operatorname{div}\left(\frac{2e^z}{-C+\sqrt{C^2+4ue^z}}\nabla u\right) = 0 \quad \text{in } Z,$$

(3.7a)
$$u|_{C_1} = \left.\frac{2\delta^4}{-C+\sqrt{C^2+4\delta^4}}\right|_{C_1}, \qquad u|_{\Gamma_Z} = 0,$$

$$\nabla u \cdot \nu|_{\partial\Omega_N \cap \partial Z} = 0,$$

$$\operatorname{div}\left(\frac{C+\sqrt{C^2+4e^w v}}{2}\nabla w\right) = 0 \quad \text{in } A,$$

(3.7b)
$$\frac{C+\sqrt{C^2+4e^w v}}{2}\nabla w \cdot \nu|_{\Gamma_A} = S\left(\frac{2e^z}{-C+\sqrt{C^2+4ue^z}}\nabla u \cdot \nu|_{\Gamma_Z}\right),$$

$$w|_{C_2} = \left.\ln\frac{C+\sqrt{C^2+4\delta^4}}{2}\right|_{C_2}, \qquad \nabla w \cdot \nu|_{\partial\Omega_N \cap \partial A} = 0;$$

$$\operatorname{div}\left(\frac{2e^w}{C+\sqrt{C^2+4e^w v}}\nabla v\right) = 0 \quad \text{in } A,$$

(3.7c)
$$v|_{C_2} = \left.\frac{2\delta^4}{C+\sqrt{C^2+4\delta^4}}\right|_{C_2}, \qquad v|_{\Gamma_A} = 0,$$

$$\nabla v \cdot \nu|_{\partial\Omega_N \cap \partial A} = 0;$$

$$\operatorname{div}\left(\frac{-C+\sqrt{C^2+4ue^z}}{2}\nabla z\right) = 0 \quad \text{in } Z,$$

(3.7d)
$$\frac{-C+\sqrt{C^2+4ue^z}}{2}\nabla z \cdot \nu|_{\Gamma_Z} = S^{-1}\left(\frac{2e^w}{C+\sqrt{C^2+4e^w v}}\nabla v \cdot \nu|_{\Gamma_A}\right),$$

$$z|_{C_1} = \left.\ln\frac{-C+\sqrt{C^2+4\delta^4}}{2}\right|_{C_1}, \qquad \nabla z \cdot \nu|_{\partial\Omega_N \cap \partial Z} = 0.$$

LEMMA 3.2. For $\delta^4 = 0$, problem (3.7) has a unique solution (u_0, v_0, w_0, z_0) that satisfies

$$u_0 = v_0 = 0, \tag{3.8a}$$

$$w_0 \in H^1(A) \cap H^2(A'), \qquad z_0 \in H^1(Z) \cap H^2(Z'),$$

$$\ln \gamma \leqq w_0, \qquad z_0 \leqq \ln (\|C\|_{L^\infty(\Omega)}) \tag{3.8b}$$

for all subdomains Z' and A' of Z and A, respectively, that do not contain critical boundary points as defined in Theorem 2.1.

Remark. Note that H^2-regularity holds in neighborhoods of Γ_Z and Γ_A whose endpoints can be considered critical boundary points for the problem (3.5), (3.6).

Proof. The maximum principle applied to problems (3.7a, c) immediately implies the estimates

$$0 \leqq u, v \leqq \frac{2\delta^4}{\gamma + \sqrt{\gamma^2 + 4\delta^4}},$$

of which (3.8a) is a direct consequence. Thus, problems (3.7b, d) for $\delta^4 = 0$ reduce to

$$\begin{aligned} &\operatorname{div}(|C|\nabla w_0) = 0 \quad \text{in } A, \\ &w_0|_{C_2} = \ln(|C|)|_{C_2}, \qquad \nabla w_0 \cdot \nu|_{\partial A \setminus C_2} = 0; \end{aligned} \tag{3.7b}_0$$

$$\begin{aligned} &\operatorname{div}(|C|\nabla z_0) = 0 \quad \text{in } Z, \\ &z_0|_{C_1} = \ln(|C|)|_{C_1}, \qquad \nabla z_0 \cdot \nu|_{\partial Z \setminus C_1} = 0. \end{aligned} \tag{3.7d}_0$$

Again applying the maximum principle, we get solutions $w_0 \in H^1(A) \cap L^\infty(A)$, $z_0 \in H^1(Z) \cap L^\infty(Z)$ and estimates (3.8b). Elliptic regularity implies $w_0 \in H^2(A')$, $z_0 \in H^2(Z')$ for A', Z', which, in addition to the requirements in the statement of the lemma, do not contain the endpoints of Γ_A and Γ_Z.

In the neighborhood of an endpoint $x_0 \in \partial\Omega$ of Γ_A we employ a local coordinate transformation as in the proof of Lemma 2.3. There it was shown that the extension of Γ_A to the exterior of Ω by reflection is smooth. The extension of w_0 to the exterior as an even function with respect to $\partial\Omega$ satisfies an elliptic equation and homogeneous Neumann conditions at the extended Γ_A. Thus, elliptic regularity results can be applied in a neighborhood of x_0, giving H^2-regularity of w_0 there. Treating endpoints of Γ_Z analogously we obtain $w_0 \in H^2(A')$, $z_0 \in H^2(Z')$ for A', Z' as in the statement of the lemma.

To facilitate the subsequent analysis we make the following additional regularity assumptions:

(A6) $w_0 \in W^{1,6}(A)$, $z_0 \in W^{1,6}(Z)$.

(A7) Let $f_0 \in L^{3/2}(\Omega)$ and $f_1, f_2 \in L^3(\Omega)$ hold.
Then the solution of

$$\begin{aligned} &\Delta w = f_0 + \operatorname{div}((f_1, f_2)) \quad \text{in } \Omega, \\ &w|_{C_1, C_2} = w_D, \qquad \nabla w \cdot \nu|_{\partial\Omega_N} = 0 \end{aligned}$$

for smooth Dirichlet data w_D is in $W^{1,3}(\Omega)$ and satisfies

$$\|w\|_{W^{1,3}(\Omega)} \leqq c_1 + c_2(\|f_0\|_{L^{3/2}(\Omega)} + \|f_1\|_{L^3(\Omega)} + \|f_2\|_{L^3(\Omega)}).$$

The validity of the above assumptions requires a certain behavior of $\partial\Omega$ close to critical boundary points. Differential equations such as those in (A7) occur when the divergence terms appearing in (3.7) are expanded, and lower-order terms are considered as data.

THEOREM 3.1. *For δ^4 small enough, problem* (3.7) *has a solution* (u, v, w, z) *in the Banach space*

$$B = [(W^{1,3}(Z) \cap H^2(Z')) \times (W^{1,3}(A) \cap H^2(A'))]^2$$

satisfying

$$\|(u-u_0, v-v_0, w-w_0, z-z_0)\|_B = O(\delta^4)$$

where Z', A' are as in Lemma 3.2.

Proof. We introduce the errors

$$\delta^4\tilde{u} = u-u_0, \quad \delta^4\tilde{v} = v-v_0, \quad \delta^4\tilde{w} = w-w_0, \quad \delta^4\tilde{z} = z-z_0.$$

By linearization we rewrite (3.7) as

$$\operatorname{div}\left(\frac{e^{z_0}}{|C|}\nabla\tilde{u} + f_1\right) = 0 \quad \text{in } Z,$$

(3.9a)
$$\tilde{u}|_{C_1} = \left.\frac{2}{-C+\sqrt{C^2+4\delta^4}}\right|_{C_1}, \qquad \tilde{u}|_{\Gamma_Z} = 0,$$

$$\nabla\tilde{u}\cdot\nu|_{\partial\Omega_N\cap\partial Z} = 0;$$

$$\operatorname{div}\left(|C|\nabla\tilde{w} + \frac{e^{w_0}}{|C|}\tilde{v}\nabla w_0 + f_2\right) = 0 \quad \text{in } A,$$

(3.9b)
$$\left(|C|\nabla\tilde{w} + \frac{e^{w_0}}{|C|}\tilde{v}\nabla w_0\right)\cdot\nu|_{\Gamma_A} = S\left(\frac{e^{z_0}}{|C|}\nabla\tilde{u}\cdot\nu|_{\Gamma_Z}\right) + f_3,$$

$$\tilde{w}|_{C_2} = \delta^{-4}\ln\left.\frac{C+\sqrt{C^2+4\delta^4}}{2C}\right|_{C_2}, \qquad \nabla\tilde{w}\cdot\nu|_{\partial\Omega_N\cap\partial A} = 0;$$

$$\operatorname{div}\left(\frac{e^{w_0}}{|C|}\nabla\tilde{v} + f_4\right) = 0 \quad \text{in } A,$$

(3.9c)
$$\tilde{v}|_{C_2} = \left.\frac{2}{C+\sqrt{C^2+4\delta^4}}\right|_{C_2}, \qquad \tilde{v}|_{\Gamma_A} = 0,$$

$$\nabla\tilde{v}\cdot\nu|_{\partial\Omega_N\cap\partial A} = 0;$$

$$\operatorname{div}\left(|C|\nabla\tilde{z} + \frac{e^{z_0}}{|C|}\tilde{u}\nabla z_0 + f_5\right) = 0 \quad \text{in } Z,$$

(3.9d)
$$\left(|C|\nabla\tilde{z} + \frac{e^{z_0}}{|C|}\tilde{u}\nabla z_0\right)\cdot\nu|_{\Gamma_Z} = S^{-1}\left(\frac{e^{w_0}}{|C|}\nabla\tilde{v}\cdot\nu|_{\Gamma_A}\right) + f_6,$$

$$\tilde{z}|_{C_1} = \delta^{-4}\ln\left.\frac{-C+\sqrt{C^2+4\delta^4}}{-2C}\right|_{C_1}, \qquad \nabla\tilde{z}\cdot\nu|_{\partial\Omega_N\cap\partial Z} = 0$$

where the operators

$$f_1, f_5 : B \to (L^3(Z)\cap H^1(Z'))^2,$$
$$f_2, f_4 : B \to (L^3(A)\cap H^1(A'))^2,$$
$$f_3 : B \to H^{1/2}(\Gamma_A), \qquad f_6 : B \to H^{1/2}(\Gamma_Z)$$

can easily be seen to be Lipschitz with $O(\delta^4)$ Lipschitz constants. The proof employs the trace theorem and the continuous imbedding $W^{1,3} \subset\subset L^\infty$. Theorem 3.1 follows by a contraction mapping argument if we can show solvability of the linear problem where the f_i are considered as inhomogeneities.

Solvability in $H^1(A)(H^1(Z))$ of (3.9a, c) is immediate. H^2-regularity away from critical boundary points follows as in the proof of Lemma 3.2. For proving $W^{1,3}$-regularity we rewrite the differential equation in (3.9a) as

$$\Delta\tilde{u} = \nabla(|C|/e^{z_0})f_1 - \operatorname{div}(|C|/e^{z_0}f_1) - |C|/e^{z_0}\nabla(e^{z_0}/|C|)\nabla\tilde{u}.$$

Using the estimates

$$\|\nabla(|C|/e^{z_0})f_1\|_{L^{3/2}(Z)} \leqq \|\nabla(|C|/e^{z_0})\|_{L^3(Z)}\|f_1\|_{L^3(Z)} \leqq c\|f_1\|_{L^3(Z)}$$

$$\||C|/e^{z_0}f_1\|_{L^3(Z)} \leqq c\|f_1\|_{L^3(Z)},$$

$$\||C|/e^{z_0}\nabla(e^{z_0}/|C|)\nabla\tilde{u}\|_{L^{3/2}(Z)} \leqq \||C|/e^{z_0}\|_{L^\infty(Z)}\|\nabla(e^{z_0}/|C|)\|_{L^6(Z)}\|\nabla\tilde{u}\|_{L^2(Z)} \leqq c\|\nabla\tilde{u}\|_{L^2(Z)}$$

where we applied assumption (A6) and the Hölder inequality, we obtain $\tilde{u} \in W^{1,3}(Z)$ and continuous dependence on the data from assumption (A7). For $\tilde{v}$ we proceed analogously.

Considering $\tilde{u}$ and $\tilde{v}$ as data in (3.9b, d) we see that these problems can obviously be solved. If we note that the trace theorem (see, e.g., [17]) guarantees a sufficiently smooth extension of the Neumann data into the interior of Z and A, respectively, the required regularity can be shown similarly to the treatment of $\tilde{u}$ and $\tilde{v}$.

Thus the proof of Theorem 3.1 is complete.

REFERENCES

[1] F. BREZZI, A. CAPELO, AND L. GASTALDI, *A singular perturbation analysis of reverse-biased semiconductor diodes*, SIAM J. Math. Anal., 20 (1989), pp. 372–387.

[2] F. BREZZI, A. CAPELO, AND L. D. MARINI, *Singular Perturbation Problems in Semiconductor Devices*, Lecture Notes in Math. 1230, J. P. Hennart, ed., Springer-Verlag, Berlin, New York, 1986, pp. 191–198.

[3] F. BREZZI AND L. GASTALDI, *Mathematical properties of one-dimensional semiconductors*, Mat. Apl. Comp., 5 (1986), pp. 123–137.

[4] L. A. CAFFARELLI AND A. FRIEDMAN, *A singular perturbation problem for semiconductors*, Boll. Un. Mat. Ital., B, 7 (1987), pp. 409–421.

[5] A. FRIEDMAN, *Variational Principles and Free-Boundary Problems*, John Wiley, New York, 1982.

[6] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1977.

[7] C. HUNT AND N. R. NASSIF, *On a variational inequality and its approximation, in the theory of semiconductors*, SIAM J. Numer. Anal., 12 (1975), pp. 938–950.

[8] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[9] J. L. LIONS, *Singular perturbations and singular layers in variational inequalities*, in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971.

[10] L. D. MARINI AND A. SAVINI, *Accurate computation of electric field in reverse-biased semiconductor devices: A mixed finite element approach*, Compel, 3 (1984), pp. 123–135.

[11] P. A. MARKOWICH, *The Stationary Semiconductor Device Equations*, Springer-Verlag, Vienna, 1986.

[12] M. S. MOCK, *Analysis of Mathematical Models of Semiconductor Devices*, Boole Press, Dublin, 1983.

[13] C. SCHMEISER, *On strongly reverse biased semiconductor diodes*, SIAM J. Appl. Math., 49 (1989), pp. 1734–1748.

[14] C. SCHMEISER AND R. WEISS, *Asymptotic analysis of singular singularly perturbed boundary value problems*, SIAM J. Math. Anal., 17 (1986), pp. 560–579.

[15] S. SELBERHERR, *Analysis and Simulation of Semiconductor Devices*, Springer-Verlag, Vienna, 1984.

[16] S. M. SZE, *Physics of Semiconductor Devices*, John Wiley, New York, 1969.

[17] G. M. TROIANIELLO, *Elliptic Differential Equations and Obstacle Problems*, Plenum Press, New York, 1987.

[18] A. B. VASILEVA AND V. F. BUTUZOV, *Singularly perturbed equations in the critical case*, Report MRC-TSR 2039, University of Wisconsin, Madison, WI, 1980.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 327-345, March 1990
© 1990 Society for Industrial and Applied Mathematics
003

STABILITY OF STEADY STATES FOR PREY-PREDATOR DIFFUSION EQUATIONS WITH HOMOGENEOUS DIRICHLET CONDITIONS*

YOSHIO YAMADA†

This paper is dedicated to Hiroshi Fujita on the occasion of his 60th *birthday.*

Abstract. This paper concerns a system of reaction-diffusion equations that describes the evolution of population densities of a prey species u and a predator species v inhabiting the same bounded domain. Under homogeneous Dirichlet boundary conditions, asymptotic stability properties of nonnegative steady states are discussed. The corresponding steady-state problem has nonnegative solutions of three different types; the trivial solution $(0, 0)$, the semitrivial solutions $(u, 0)$, $(0, v)$ with u, v positive, and a positive solution (u, v) with both components positive. Stability properties of the trivial and semitrivial solutions are determined completely. The stability and uniqueness of positive solutions are also studied. This method is based on spectral analysis, comparison principle, and bifurcation theory.

Key words. reaction-diffusion equations, positive steady state, asymptotic stability, bifurcation, prey-predator system

AMS(MOS) subject classifications. primary 35K55; secondary 35B32, 35B40

1. Introduction. In this paper we study the following system of reaction-diffusion equations:

$$\begin{aligned} u_t &= d_1\Delta u + au + uf(u, v), \\ v_t &= d_2\Delta v + bv + vg(u, v), \end{aligned} \tag{1.1}$$

for $x = (x_1, x_2, \cdots, x_N) \in \Omega$ and $t > 0$, where Δ is the Laplacian, and Ω is a bounded domain in R^N with smooth boundary Γ. In (1.1), d_1 and d_2 are positive diffusion constants, a and b are some constants, and the interaction between u and v is determined by f and g. These equations are supplemented by homogeneous Dirichlet boundary conditions

$$u = v = 0 \quad \text{for } (x, t) \in \Gamma \times (0, \infty). \tag{1.2}$$

Systems such as (1.1) arise in mathematical ecology and describe the evolution of population densities of two interacting species that inhabit the same region Ω undergoing simple diffusion. We study (1.1) as a model of prey-predator systems. Let u and v be population densities of a prey and a predator species, respectively. The constant a in (1.1) represents the birth rate of u and is assumed to be positive. Similarly, the constant b represents the birth rate of v, but we do not require its positivity. The functions f and g, reflecting the dynamics of the prey-predator interaction between u and v, are assumed to fulfill the following conditions:

(A.1) $f(u, v)$ is a C^1-function of $(u, v) \in [0, \infty) \times [0, \infty)$ such that $f(0, 0) = 0$ and $f_v(u, v) < 0$ for all $(u, v) \in [0, \infty) \times [0, \infty)$.

(A.2) $g(u, v)$ is a C^1-function of $(u, v) \in [0, \infty) \times [0, \infty)$ such that $g(0, 0) = 0$ and $g_u(u, v) > 0$ for all $(u, v) \in [0, \infty) \times [0, \infty)$.

* Received by the editors August 12, 1988; accepted for publication (in revised form) April 21, 1989.
† Department of Mathematics, Waseda University, Ohkubo 3-4-1, Shinjuku, Tokyo, 160 Japan.

Moreover, we impose the following self-limiting assumptions on f and g:

(A.3) $f(u, 0)$ is strictly monotone decreasing in $u \geqq 0$ and $\lim_{u\to\infty} f(u, 0) = -\infty$.

(A.4) $g(0, v)$ is strictly monotone decreasing in $v \geqq 0$ and $\lim_{v\to\infty} g(u, v) = -\infty$ for each $u \geqq 0$.

According to (A.3) and (A.4), each species can control its growth rate in the absence of the other species. Boundary condition (1.2) means that the habitat Ω is surrounded by hostile environment for both species.

Our purpose is to study the asymptotic behavior of nonnegative solutions for (1.1)-(1.2). In connection with this study, we are interested in finding all nonnegative steady-state solutions for (1.1)-(1.2) and deciding their stability. Related problems have been discussed by many authors; for prey-predator systems with Dirichlet conditions, see, e.g., [1]-[5], [8], [9], [14]-[16]. Especially, Blat and Brown [1] have studied nonnegative steady-state solutions for (1.1)-(1.2) in the case where

$$f(u, v) = -a_1 u - a_2 v, \qquad g(u, v) = b_1 u - b_2 v, \tag{1.3}$$

for some positive constants a_1, a_2, b_1, and b_2. By making use of decoupling and global bifurcation techniques, they have constructed nonnegative steady-state solutions, including positive ones. Recently, their existence results have been sharpened by Dancer [8], [9] with the use of degree theory. Indeed, necessary and sufficient conditions are established for the existence of positive steady-state solutions. (See also Li [16].)

However, it seems that stability properties of steady states have not yet been completely understood. In this direction, we refer the reader to the work of Conway, Gardner, and Smoller [4], [5], who have discussed the change of stability of nonnegative steady states for similar problems in the case $N = 1$.

We will indicate our main results on the stability of nonnegative steady states as well as on their structure. For the sake of simplicity, we choose f and g of form (1.3) and take a and b as bifurcation parameters. As in Fig. 1, the (a, b) parameter space is divided into four regions I, II, III, and IV defined by positive constants a^*, b^* and monotone C^1-functions $\bar{a}$, $\bar{b}$. Curves C_1 and C_2 are defined by $b = \bar{b}(a)$ and $a = \bar{a}(b)$, respectively. Besides the trivial solution $(0, 0)$, the stationary problem for (1.1)-(1.2) has two semitrivial solutions as nonnegative steady states: $(u^*, 0)$ for $a > a^*$ and $(0, v^*)$ for $b > b^*$. When (a, b) lies in I, $(0, 0)$ is a global attractor for (1.1)-(1.2); that is, all nonnegative solutions of (1.1)-(1.2) converge to $(0, 0)$ as $t \to \infty$. As a increases across a^* for each fixed b, then $(u^*, 0)$ bifurcates from $(0, 0)$. For $(a, b) \in$ II, $(u^*, 0)$ is a global attractor to (1.1)-(1.2) with nonnegative initial data, whereas $(0, 0)$ loses its stability. Another semitrivial solution $(0, v^*)$ existing for $b > b^*$ has similar stability properties when $(a, b) \in$ III. As a result, $(0, 0)$, $(u^*, 0)$, and $(0, v^*)$, respectively, possess I, II, and III as their global stability regions, so that there are no positive steady states for $(a, b) \in \text{I} \cup \text{II} \cup \text{III}$. When (a, b) lies in IV, these trivial and semitrivial solutions become unstable and a positive steady-state solution $(\bar{u}, \bar{v})$ appears as a secondary bifurcation from $(u^*, 0)$ or $(0, v^*)$. Especially, $(\bar{u}, \bar{v})$ is locally stable for $(a, b) \in$ IV restricted in a neighborhood of $C_1 \cup C_2$.

The content of this paper is as follows. In § 2 we collect some preliminary results about asymptotic and stability properties for related single reaction-diffusion equations. In § 3 we carry out spectral analysis for trivial and semitrivial steady-state solutions of (1.1)-(1.2). Section 4 is devoted to the study of global attractivity of trivial and semitrivial solutions, where the comparison principle is a basic tool. By the local bifurcation theory due to Crandall and Rabinowitz [6], [7], it is shown in § 5 that positive steady-state solutions bifurcate from two semitrivial ones and that they are

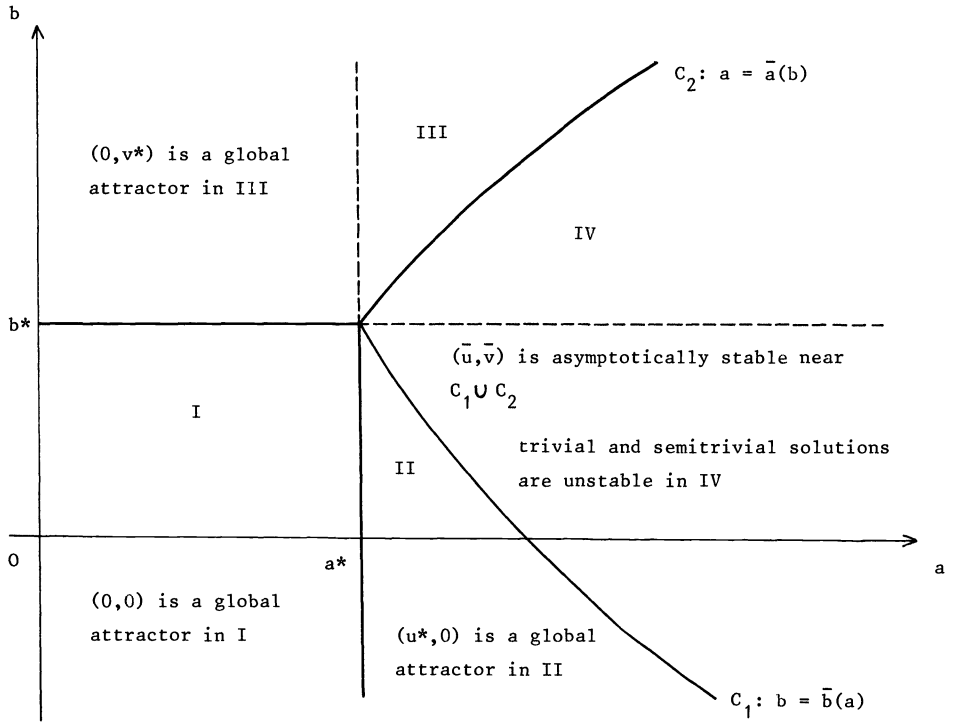


FIG. 1

locally stable. According to global bifurcation theory, these bifurcating solutions are connected to each other by a branch of positive steady-state solutions. In § 6 we carry out bifurcation and stability analysis near $(a, b) = (a^*, b^*)$ to get more information about the branch of positive solutions.

The analysis in the present paper concentrates on reaction-diffusion systems of prey-predator type. However, most of the methods developed in §§ 3, 5, and 6 are also valid for studying reaction-diffusion systems of competition type.

Notation. The usual norms of the spaces $L^p(\Omega)$ for $1 \leqq p < \infty$ and $C(\bar{\Omega})$ are defined by

$$\|u\|_p^p = \int_\Omega |u(x)|^p \, dx, \qquad \|u\|_\infty = \max_{x \in \bar{\Omega}} |u(x)|.$$

In particular, we denote by $(\cdot, \cdot)_2$ the inner product of $L^2(\Omega)$. For any integer k, let $W^{k,p}(\Omega)$ be the Sobolev space of functions $u: \Omega \to R$ such that u and its distributional derivatives up to order k belong to $L^p(\Omega)$. The completion of $C_0^\infty(\Omega)$ with respect to the $W^{k,p}(\Omega)$-norm is denoted by $\mathring{W}^{k,p}(\Omega)$.

2. Some preliminaries for single reaction-diffusion equations. In this section we collect some results about single reaction-diffusion equations that are required later.

Consider the following initial boundary value problem:

$$\begin{aligned} &w_t = d\Delta w + \alpha(x)w + (c + h(w))w \quad \text{in } \Omega \times (0, \infty), \\ &w = 0 \qquad\qquad\qquad\qquad\qquad\quad \text{on } \Gamma \times (0, \infty), \end{aligned} \tag{2.1}$$

where d is a positive constant, α is a $C^{\mu}(\bar{\Omega})$-function with $\mu > 0$, and c is a parameter moving over $(-\infty, \infty)$. We assume that

(H) $h:[0,\infty)\to(-\infty,0]$ is a strictly monotone decreasing function of class $C^1([0,\infty))$ such that $h(0)=0$ and $\lim_{w\to\infty} h(w)=-\infty$.

Let $p > N$ and set $X = L^p(\Omega)$. We define a closed linear operator A in X with domain $D(A)$ by

$$Au = -d\Delta w \quad \text{for } w \in D(A) \equiv W^{2,p}(\Omega) \cap \mathring{W}^{1,p}(\Omega).$$

It is well known that $-A$ generates an analytic semigroup $\{e^{-tA}\}_{t\geqq 0}$ and that the solution w of (2.1) with $w(0) = w_0$ satisfies

$$w(t) = e^{-tA}w_0 + \int_0^t e^{-(t-s)A}\tilde{h}(w(s))\,ds, \tag{2.2}$$

where $\tilde{h}(w) = (\alpha + c + h(w))w$.

Let w_0 be any nonnegative function of class $C(\bar{\Omega})$ and let $w(x, t; w_0)$ denote the solution of (2.1) with $w(0) = w_0$. By virtue of (H), we can choose $M > 0$ such that $h(M) + c + \|\alpha\|_{\infty} = 0$. Then the comparison theorem (see, e.g., Protter and Weinberger [17]) yields an a priori estimate

$$0 \leqq w(x, t; w_0) \leqq \max\{\|w_0\|_{\infty}, M\}, \tag{2.3}$$

which assures that the solution w of (2.1) exists globally in time. Moreover, using (2.3) and the smoothing effect of parabolic equations, we can see from (2.2) that, for any $0 \leqq \alpha < 1$, the solution orbit $\{w(t; w_0);\ t \geqq 1\}$ is relatively compact in $D(A^{\alpha})$ equipped with graph norm of A^{α} (see Henry [10, Thm. 3.3.6]). Here we observe that

$$D(A^{\alpha}) \subset C^1(\bar{\Omega}) \quad \text{if } 1 + N/p < 2\alpha, \tag{2.4}$$

where injection is continuous and compact (see, e.g., [10, Thms. 1.4.8 and 1.6.1]). In what follows, we fix α satisfying $(p+N)/2p < \alpha < 1$. Define the ω-limit set $\omega(w_0)$ associated with the solution orbit $\{w(t; w_0);\ t \geqq 0\}$ by

$$\begin{aligned}\omega(w_0) = \{w^*;\ &\text{there exists a sequence } \{t_n\}\uparrow\infty \\ &\text{such that } w(t_n; w_0) \to w^* \text{ in } D(A^{\alpha})\};\end{aligned} \tag{2.5}$$

then it is nonempty, connected, and invariant [10, Thm. 4.3.3]. Furthermore, (2.1) is gradient-like with respect to the functional

$$E(w) = \int_{\Omega} \{d|\nabla w(x)|^2/2 - (\alpha(x) + c)w(x)^2/2 - H(w(x))\}\,dx,$$

where $H'(w) = wh(w)$, so that E is a Lyapunov function on $D(A^{\alpha})$, that is, $dE(w(t; w_0))/dt \leqq 0$. Therefore,

$$\omega(w_0) \subset \{w^* \in D(A^{\alpha});\ w^* \text{ is a nonnegative steady-state solution of (2.1)}\} \tag{2.6}$$

[10, Thm. 4.3.4]; so that any $w^* \in \omega(w_0)$ satisfies

$$\begin{aligned} &-d\Delta w - \alpha w = (c + h(w))w \quad \text{and} \quad w \geqq 0 \quad \text{in } \Omega, \\ &w = 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{on } \Gamma. \end{aligned} \tag{2.7}$$

Let ζ_0 be the least eigenvalue for

$$-d\Delta w - \alpha w = \zeta w \quad \text{in } \Omega \quad \text{and} \quad w = 0 \quad \text{on } \Gamma. \tag{2.8}$$

It is well known that, if $c \leqq \zeta_0$, then $w \equiv 0$ is the only solution of (2.7), whereas if $c > \zeta_0$, then (2.7) has a unique positive solution w^* (see, e.g., Blat and Brown [1]). Thus we are able to show the following proposition in the standard manner with use of the comparison technique (see, e.g., [5, Prop. 2.1]).

PROPOSITION 2.1. *Let w be the solution of* (2.1) *with $w(\cdot\,, 0) = w_0$, where $w_0 (\not\equiv 0)$ is a nonnegative function of class $C(\bar{\Omega})$.*

(i) *If $c \leqq \zeta_0$, then $\lim_{t\to\infty} w(\cdot\,, t) = 0$ in $C(\bar{\Omega})$.*

(ii) *If $c > \zeta_0$, then $\lim_{t\to\infty} w(\cdot\,, t) = w^*(c)$ in $C(\bar{\Omega})$, where $w^*(\cdot\,; c)$ is a unique positive solution of* (2.7).

We will investigate some properties related to $w^*(c)$ for $c > \zeta_0$. The linearized operator $L(c): X \to X$ of (2.7) about $w^*(c)$ is given by

$$L(c)w = Aw - \alpha w - cw - (wh(w))'(w^*(c))w \quad \text{with } D(L(c)) = D(A),$$

where $(wh(w))' = d(wh(w))/dw$. As is well known, the spectrum $\sigma(L(c))$ of $L(c)$ consists of eigenvalues that lie on the real axis. Furthermore, we can prove the following lemma.

LEMMA 2.2. *For each $c > \zeta_0$, all eigenvalues of $L(c)$ are positive.*

Proof. It suffices to follow the arguments used by Ito [11, Lemma A.1] (see also Blat and Brown [2, Lemma 2.1]). □

LEMMA 2.3. *For each $c > \zeta_0$, $L(c)$ has a bounded inverse $L(c)^{-1}$. Moreover,*

$$L(c)^{-1} f > 0 \quad \text{in } \Omega$$

for all f $(\not\equiv 0) \in X$ satisfying $f \geqq 0$ almost everywhere in Ω.

Proof. If follows from Lemma 2.2 that $L(c)$ is invertible in X. For each nonnegative $f \in X$, put $w = L(c)^{-1} f$ and decompose it as $w = w^+ - w^-$, where $w^+ = \max\{w, 0\}$ and $w^- = -\min\{w, 0\}$. Since w satisfies

$$-d\Delta w - \{\alpha + c + (wh(w))'(w^*(c))\}w = f \quad \text{in } \Omega \tag{2.9}$$

with $w = 0$ on Γ, multiplying (2.9) by $-w^-$ and integrating the resulting expression over Ω, we can derive

$$0 \geqq -\int_\Omega f w^- \, dx = \int_\Omega [d|\nabla w^-|^2 - \{\alpha + c + (wh(w))'(w^*(c))\}(w^-)^2] \, dx$$

$$\geqq \nu_0 \int_\Omega (w^-)^2 \, dx,$$

where ν_0 is the least eigenvalue of $L(c)$, which is positive by Lemma 2.2. Thus, $w^- \equiv 0$; so that $w \geqq 0$ almost everywhere in Ω. The strict positivity of w in Ω follows from the strong maximum principle. □

Finally, we state the dependence of $w^*(c)$ on c.

LEMMA 2.4. (i) $\operatorname{Lim}_{c\to\zeta_0} w^*(c) = 0$ *uniformly in $\bar{\Omega}$.*

(ii) *The mapping $c \to w^*(c)$ is of class $C^1((\zeta_0, \infty); D(A))$ and the Fréchet derivative of w^* with respect to c, which is denoted by w_c^*, satisfy*

$$w_c^*(c) > 0 \quad \text{in } \Omega \quad \text{for } c > \zeta_0.$$

For the proof, see [2, Lemma 2.2] or [11, Lemma A.3].

3. Spectral analysis. We first give existence and regularity results of global solutions for (1.1)-(1.2) with nonnegative initial data.

For $p > n$, set $\mathbf{X} = \{L^p(\Omega)\}^2$, $\mathbf{Y} = \{W^{2,p}(\Omega) \cap \mathring{W}^{1,p}(\Omega)\}^2$ and define a closed linear operator $\mathbf{A}$ in $\mathbf{X}$ by

$$\mathbf{A}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} A_1 u \\ A_2 v \end{pmatrix} \quad \text{for} \begin{pmatrix} u \\ v \end{pmatrix} \in D(\mathbf{A}) \equiv \mathbf{Y},$$

where $A_1 u = -d_1 \Delta u$ and $A_2 v = -d_2 \Delta v$. Since $-\mathbf{A}$ generates an analytic semigroup $\{\exp(-t\mathbf{A})\}_{t \geqq 0}$ in $\mathbf{X}$, the initial value problem for (1.1), (1.2) can be treated as that for an abstract evolution equation (see § 2).

Then the global solvability theorem reads as follows.

PROPOSITION 3.1. *Let u_0, v_0 be nonnegative functions of class $C(\bar{\Omega})$. Then there exists a unique solution (u, v) of* (1.1), (1.2) *with $(u(0), v(0)) = (u_0, v_0)$ in the class $C([0, \infty); \mathbf{X}) \cap C^1((0, \infty); D(\mathbf{A}))$. Moreover,*

$$0 \leqq u(x, t) \leqq m_1 \quad \text{and} \quad 0 \leqq v(x, t) \leqq m_2 \tag{3.1}$$

for $(x, t) \in \Omega \times [0, \infty)$ with some positive constants m_1, m_2.

Proof. Since the local solvability and uniqueness can be derived in the standard manner, to obtain the global solvability it suffices to show (3.1) (see, e.g., Rothe [19, Thm. 1]). In view of the forms in (1.1), it is easy to see from the comparison theorem that

$$u \geqq 0 \quad \text{and} \quad v \geqq 0 \tag{3.2}$$

as long as the solution (u, v) exists. Hence, by (A.1),

$$u_t \leqq d_1 \Delta u + u(a + f(u, 0)). \tag{3.3}$$

Since we can choose $M_a > 0$ satisfying $a + f(M_a, 0) = 0$ by (A.3), application of the comparison theorem to (3.3) yields

$$u \leqq \max\{\|u_0\|_\infty, M_a\} \equiv m_1. \tag{3.4}$$

Therefore, because of (3.4) and (A.2), we can show

$$v_t \leqq d_2 \Delta v + v(b + g(m_1, v)). \tag{3.5}$$

As in the derivation of (3.4), it follows from (3.5) that

$$v \leqq \max\{\|v_0\|_\infty, N_b\} \equiv m_2, \tag{3.6}$$

where N_b is a positive constant satisfying $b + g(m_1, N_b) = 0$ (use (A.4)). Thus (3.1) is derived from (3.2), (3.4), and (3.6). □

Usually, the asymptotic behavior of global solutions is closely related to the stability analysis of the corresponding stationary problem. So we consider

$$-d_1 \Delta u - au = uf(u, v) \quad \text{in } \Omega, \tag{3.7}$$

$$-d_2 \Delta v - bv = vg(u, v) \quad \text{in } \Omega, \tag{3.8}$$

$$u = v = 0 \quad \text{on } \Gamma, \tag{3.9}$$

with additional condition

$$u \geqq 0 \quad \text{and} \quad v \geqq 0 \quad \text{in } \Omega. \tag{3.10}$$

Besides the *trivial solution* $(0, 0)$, the stationary problem above may have solutions of two different types; *semitrivial solutions* $(u, 0)$, $(0, v)$ with u and v positive and *positive solution* (u, v) with both components positive. The existence of semitrivial

solutions follows from Proposition 2.1. Indeed, if we denote by λ_0 (>0) the least eigenvalue of

$$-\Delta u = \lambda u \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \Gamma, \tag{3.11}$$

then Proposition 2.1, together with (A.1) and (A.3), yields the existence of a (unique) positive function $u^*(a)$ satisfying

$$-d_1\Delta u^* - au^* = u^* f(u^*, 0) \quad \text{in } \Omega, \qquad u^* = 0 \quad \text{on } \Gamma, \tag{3.12}$$

for $a > d_1\lambda_0$; so that $(u^*(a), 0)$ clearly satisfies (3.7)-(3.10). Similarly, there is another semitrivial solution $(0, v^*(b))$ for $b > d_2\lambda_0$, where $v^*(b)$ is the unique positive solution of

$$-d_2\Delta v^* - bv^* = v^* g(0, v^*) \quad \text{in } \Omega, \qquad v^* = 0 \quad \text{on } \Gamma. \tag{3.13}$$

We now investigate the (local) stability of these trivial and semitrivial solutions by examining the spectrum of the corresponding linearized operator. Recall that any solution $(\bar{u}, \bar{v})$ of (3.7)-(3.9) is said to be *asymptotically stable* if the spectrum of the linearized operator of (3.7)-(3.9) at $(\bar{u}, \bar{v})$ lies in the right-hand side of the imaginary axis. If there are some points in the spectrum with negative real parts, we say that $(\bar{u}, \bar{v})$ is *unstable*. (For details, see Kielhöfer [13, Thms. 4.1, 4.2] or Henry [10, Thms. 5.1.1, 5.1.3].)

The linearized operator of (3.7)-(3.9) at $(0, 0)$ is given by

$$\mathbf{L}_0(a, b) = \begin{pmatrix} A_1 - a & 0 \\ 0 & A_2 - b \end{pmatrix}, \tag{3.14}$$

whose spectrum consists of eigenvalues. Clearly, we have Theorem 3.2.

THEOREM 3.2. *Set*

$$a^* = d_1\lambda_0 \quad \textit{and} \quad b^* = d_2\lambda_0. \tag{3.15}$$

The trivial solution is asymptotically stable if $a < a^*$ and $b < b^*$ and unstable if $a > a^*$ or $b > b^*$.

We proceed to the spectral analysis of the semitrivial solutions. The linearized operator of (3.7)-(3.9) at $(u^*(a), 0)$ produces the closed operator $\mathbf{L}_1(a, b)$ in $\mathbf{X}$ given by

$$\mathbf{L}_1(a, b) = \begin{pmatrix} L_1(a) & -u^*(a) f_v(u^*(a), 0) \\ 0 & L_2(a, b) \end{pmatrix} \tag{3.16}$$

with $D(\mathbf{L}_1) = \mathbf{Y}$, where $L_1(a)u = A_1 u - au - (uf)_u(u^*(a), 0)u$ and $L_2(a, b)v = A_2 v - bv - g(u^*(a), 0)v$. By the Riesz-Schauder theory, the spectrum $\sigma(\mathbf{L}_1(a, b))$ of $\mathbf{L}_1(a, b)$ consists of real eigenvalues and

$$\sigma(\mathbf{L}_1(a, b)) = \sigma(L_1(a)) \cup \sigma(L_2(a, b))$$

(cf. [11, Lemma 3.5] or [5, Thm. 2.7]). By Lemma 2.2, $\sigma(L_1(a))$ lies on the positive real axis. Moreover, $\sigma(L_2(a, b))$ lies on the real axis and the least eigenvalue $\nu_2(a, b)$ is characterized as

$$\nu_2(a, b) = \bar{b}(a) - b, \tag{3.17}$$

with

$$\bar{b}(a) = \inf\{d_2\|\nabla v\|_2^2 - (g(u^*(a), 0)v, v)_2;\ v \in \mathring{W}^{1,2}(\Omega),\ \|v\|_2 = 1\}. \tag{3.18}$$

Therefore, the stability of $(u^*(a), 0)$ is stated as follows.

THEOREM 3.3. *Define* (a^*, b^*) *by* (3.15) *and* $\bar{b}$ *by* (3.18). *Let* $a \geqq a^*$. *Then* $(u^*(a), 0)$ *is asymptotically stable if* $\bar{b}(a) > b$ *and unstable if* $\bar{b}(a) < b$.

For later use we will collect some properties of $\bar{b}$.

LEMMA 3.4. *The function* $\bar{b}$ *defined by* (3.18) *satisfies the following*:

(i) $\bar{b} \in C([a^*, \infty))$ *and* $\bar{b}(a^*) = b^*$.

(ii) $\bar{b} \in C^1((a^*, \infty))$ *and* $\bar{b}'(a) < 0$.

(iii) If $f_u(0, 0) < 0$, *then* $\lim_{a \to a^*} \bar{b}'(a) = g_u(0, 0)/f_u(0, 0) < 0$.

Proof. We follow the idea used by Ito [11, Lemma A.4] to prove (i) and (ii). Since $\bar{b}(a)$ is the least eigenvalue of

$$-d_2 \Delta v - g(u^*(a), 0)v = \mu v \quad \text{in } \Omega, \qquad v = 0 \quad \text{on } \Gamma,$$

we may take the corresponding eigenfunction $\psi(a)$ such that $\|\psi(a)\|_2 = 1$ and $\psi(a) > 0$ in Ω. The infimum in (3.18) is attained by $\psi(a)$, so that

$$\begin{aligned} \bar{b}(a+h) &= d_2 \|\nabla \psi(a+h)\|_2^2 - (g(u^*(a+h), 0)\psi(a+h), \psi(a+h))_2 \\ &\leqq d_2 \|\nabla \psi(a)\|_2^2 - (g(u^*(a+h), 0)\psi(a), \psi(a))_2 \\ &= \bar{b}(a) - (\{g(u^*(a+h), 0) - g(u^*(a), 0)\}\psi(a), \psi(a))_2. \end{aligned} \tag{3.19}$$

The similar inequality is valid if a and $a+h$ are exchanged. Hence

$$|\bar{b}(a+h) - \bar{b}(a)| \leqq \|g(u^*(a+h), 0) - g(u^*(a), 0)\|_\infty. \tag{3.20}$$

Thus the assertion of (i) follows from (3.20) and Lemma 2.4. To prove (ii), we observe that (3.20), together with (ii) of Lemma 2.4, implies the local Lipschitz continuity of $\bar{b}$ on (a^*, ∞). Hence making use of (3.19), divide $\bar{b}(a+h) - \bar{b}(a)$ by $h > 0$ ($h < 0$) and let $h \to 0$; then

$$\bar{b}'(a) = -\int_\Omega g_u(u^*(a), 0) u_a^*(a) \psi(a)^2 \, dx \tag{3.21}$$

for almost every $a \in (a^*, \infty)$. In (3.21), $a \to u^*(a)$ is continuously differentiable in $C(\bar{\Omega})$ by Lemma 2.4 and $a \to \psi(a)$ is continuous in $L^2(\Omega)$ by the perturbation result of Kato [12, Chap. 4, § 5]. Thus the right-hand side of (3.21) is continuous in a, so that (3.21) holds true for all $a \in (a^*, \infty)$. In view of (A.2) and Lemma 2.4, the negativity of $\bar{b}'$ easily follows from (3.21).

We prove (iii) with use of the identity (3.21). Let φ_0 be the eigenfunction of (3.11) corresponding to the least eigenvalue λ_0; so that we can take φ_0 satisfying $\varphi_0 > 0$ in Ω and $\|\varphi_0\|_2 = 1$. By (i) of Lemma 2.4, it is possible to show that

$$\lim_{a \to a^*} g_u(u^*(a), 0) = g_u(0, 0) \quad \text{in } C(\bar{\Omega}). \tag{3.22}$$

Moreover, by the result of Kato [12],

$$\lim_{a \to a^*} \psi(a) = \varphi_0 \quad \text{in } L^2(\Omega). \tag{3.23}$$

It remains to derive the dependence of u_a^* on a. Since $u^*(a)$ is the unique nontrivial solution of (3.12) for $a > a^*$, we make use of the local bifurcation theory of Crandall and Rabinowitz [7, Lemma 1.1] to get an expression of $u^*(a)$ near $a = a^*$. Set $X = L^p(\Omega)$ and $Y = W^{2,p}(\Omega) \cap \mathring{W}^{1,p}(\Omega)$. Define the operator $F: Y \times R \to X$ by

$$F(u; a) = A_1 u - au - uf(u, 0).$$

Clearly, $F(0; a)=0$ and $F_u(0; a^*)=A_1-a^*$; so that $\dim N(A_1-a^*)=$ codim $R(A_1-a^*)=1$, where the null space and range space of a linear operator L are denoted by $N(L)$ and $R(L)$. Moreover,

$$F_{ua}(0; a^*)\varphi_0=-\varphi_0\notin R(A_1-a^*).$$

Therefore, applying Lemma 1.1 of [7], we obtain functions $(w(s), a(s))\in C^1([-s_0, s_0]; X\times R)$, where s_0 is a sufficiently small number, with the following properties: (a) $w(0)=0$ and $a(0)=a^*$; (b) $F(u(s), a(s))=0$, where $u(s)=s(\varphi_0+w(s))$ and $w(s)\in R(A_1-a^*)\cap Y$. The uniqueness of nontrivial solutions for (3.12) near $a=a^*$ (see [7, Lemma 1.1]) implies

$$u^*(a(s))=s(\varphi_0+w(s)). \tag{3.24}$$

Now observe that

$$\begin{aligned}0&=(F(s(\varphi_0+w(s)); a(s)), \varphi_0)_2\\&=-s(a(s)-a^*)-s((\varphi_0+w(s))f(s(\varphi_0+w(s)), 0), \varphi_0)_2\end{aligned}$$

because $\|\varphi_0\|_2=1$ and $w(s)\in R(A_1-a^*)$. Hence

$$a(s)-a^*=-((\varphi_0+w(s))f(s(\varphi_0+w(s)), 0), \varphi_0)_2;$$

so

$$\frac{da}{ds}(0)=-f_u(0, 0)\|\varphi_0\|_3^3. \tag{3.25}$$

Since it follows from (3.24) and (3.25) that

$$u_a^*(a^*)=\varphi_0/a_s(0)=-\varphi_0(f_u(0, 0)\|\varphi_0\|_3^3)^{-1}, \tag{3.26}$$

identity (3.21), together with (3.22), (3.23), and (3.26), gives

$$\lim_{a\to a^*}\bar{b}'(a)=\frac{g_u(0, 0)}{f_u(0, 0)}. \qquad \square$$

The stability analysis for $(0, v^*(b))$ can be carried out in the same way as that for $(u^*(a), 0)$. The linearized operator $\mathbf{L}_2(a, b)$: $\mathbf{X}\to\mathbf{X}$ of (3.7)-(3.9) at $(0, v^*(b))$ is given by

$$\mathbf{L}_2(a, b)=\begin{pmatrix}L_3(a, b) & 0\\ -v^*(b)g_u(0, v^*(b)) & L_4(b)\end{pmatrix}$$

with $D(\mathbf{L}_2(a, b))=\mathbf{Y}$, where $L_3(a, b)u=A_1u-au-f(0, v^*(b))u$ and $L_4(b)v=A_2v-bv-(vg)_v(0, v^*(b))v$. The spectrum

$$\sigma(\mathbf{L}_2(a, b))=\sigma(L_3(a, b))\cup\sigma(L_4(b))$$

is composed of only real eigenvalues. By Lemma 2.2, the least eigenvalue of $L_4(b)$ is positive. To characterize the least eigenvalue $\nu_3(a, b)$ of $L_3(a, b)$, we introduce the following function (cf. (3.18)):

$$\bar{a}(b)=\inf\{d_1\|\nabla u\|_2^2-(f(0, v^*(b))u, u)_2;\ u\in\mathring{W}^{1,2}(\Omega), \|u\|_2=1\}. \tag{3.27}$$

Since $\nu_3(a, b)=\bar{a}(b)-a$, the stability of $(0, v^*(b))$ reads as follows.

THEOREM 3.5. *Let $b\geqq b^*$. If $\bar{a}$ is defined by* (3.27), *then $(0, v^*(b))$ is asymptotically stable if $a<\bar{a}(b)$ and unstable if $a>\bar{a}(b)$.*

Finally, we state some properties of the function $\bar{a}$, which can be shown in the same way as Lemma 3.4.

LEMMA 3.6. *The function* $\bar{a}$ *defined by* (3.27) *satisfies*

(i) $\bar{a} \in C([b^*, \infty))$ *and* $\bar{a}(b^*) = a^*$;

(ii) $\bar{a} \in C^1((b^*, \infty))$ *and* $\bar{a}'(b) > 0$;

(iii) *If* $g_v(0, 0) < 0$, *then*

$$\lim_{b \to b^*} \bar{a}'(b) = f_v(0, 0)/g_v(0, 0) > 0.$$

4. Stability analysis via the comparison principle. In § 3 we discussed the linearized stability of trivial and semitrivial steady-state solutions. More information about their stability properties can be derived via the comparison principle.

Though Theorem 3.2 means merely the local stability of $(0, 0)$, its global attractivity holds in the following sense.

THEOREM 4.1. *If* $a \leqq a^*$ *and* $b < b^*$, *then every nonnegative solution of* (1.1), (1.2) *converges to* $(0, 0)$ *uniformly in* $\bar{\Omega}$ *as* $t \to \infty$.

Proof. By the nonnegativity of u and (A.1),

$$u_t \leqq d_1 \Delta u + au + uf(u, 0) \quad \text{in } \Omega \times (0, \infty).$$

Let U be the solution of

$$\begin{aligned} &U_t = d_1 \Delta U + aU + Uf(U, 0) && \text{in } \Omega \times (0, \infty), \\ &U = 0 && \text{on } \Gamma \times (0, \infty), \\ &U(\cdot, 0) = u(\cdot, 0) \geqq 0 && \text{in } \Omega. \end{aligned} \tag{4.1}$$

The comparison theorem for parabolic equations implies $0 \leqq u \leqq U$ and Proposition 2.1 ensures that $U(\cdot, 0) \to 0$ uniformly in $\bar{\Omega}$ as $t \to \infty$. Therefore,

$$\lim_{t \to \infty} u(\cdot, t) = 0 \quad \text{uniformly in } x \in \bar{\Omega}. \tag{4.2}$$

Because of (4.2) we can show that, for any $\varepsilon > 0$, there exists T_ε such that

$$v_t \leqq d_2 \Delta v + (b + \varepsilon)v + vg(0, v) \quad \text{in } \Omega \times [T_\varepsilon, \infty).$$

Since $b + \varepsilon \leqq b^*$ for sufficiently small ε, Proposition 2.1, together with the comparison theorem, enables us to show that

$$\lim_{t \to \infty} v(\cdot, t) = 0 \quad \text{uniformly in } x \in \bar{\Omega}, \tag{4.3}$$

as in the derivation of (4.2). Thus (4.2) and (4.3) yield the assertion. □

Before studying the global attractivity of $(u^*(a), 0)$ or $(0, v^*(b))$, we put some additional conditions on f and g:

$$f(0, v) \geqq f(u, v) \quad \text{for all } u, v \geqq 0, \tag{A.5}$$

$$g(u, 0) \geqq g(u, v) \quad \text{for all } u, v \geqq 0. \tag{A.6}$$

THEOREM 4.2. *In addition to* (A.1)-(A.4), *assume* (A.6). *If* $a > a^*$ *and* $\bar{b}(a) > b$, *then every nonnegative solution* (u, v) *of* (1.1), (1.2) *with* $u(\cdot, 0) \not\equiv 0$ *satisfies*

$$\lim_{t \to \infty} (u(\cdot, t), v(\cdot, t)) = (u^*(a), 0) \quad \textit{uniformly in } \bar{\Omega}. \tag{4.4}$$

Proof. We employ the same method, based on the comparison theorem, as that used by Conway, Gardner, and Smoller [5, Thm. 2.3] (see also [4]). Note that the solution U of (4.1) satisfies

$$\lim_{t\to\infty} U(\cdot, t) = u^*(a) \quad \text{uniformly in } \bar{\Omega}$$

by Proposition 2.1. Therefore, in view of $u \leqq U$, we can show that

$$\limsup_{t\to\infty} u(\cdot, t) \leqq u^*(a) \quad \text{uniformly in } \bar{\Omega}. \tag{4.5}$$

This fact, together with (A.2) and (A.6), assures that, for any $\varepsilon > 0$, there is T_ε such that

$$\begin{aligned} v_t &\leqq d_2\Delta v + (b+\varepsilon)v + vg(u^*(a), v) \\ &\leqq d_2\Delta v + (b+\varepsilon)v + vg(u^*(a), 0) \end{aligned}$$

in $\Omega \times [T_\varepsilon, \infty)$. Here we recall that the least eigenvalue of $-d_2\Delta - g(u^*(a), 0)$ with homogeneous Dirichlet condition is $\bar{b}(a)$ (see (3.18)). Since $b+\varepsilon < \bar{b}(a)$ if $\varepsilon > 0$ is sufficiently small, it follows from the comparison theorem that

$$\lim_{t\to\infty} v(\cdot, t) = 0 \quad \text{uniformly in } \bar{\Omega}. \tag{4.6}$$

Having established (4.6), we return to the first equation of (1.1). For any $\varepsilon > 0$,

$$u_t \geqq d_1\Delta u + (a-\varepsilon)u + uf(u, 0) \quad \text{in } \Omega \times [T'_\varepsilon, \infty)$$

with some $T'_\varepsilon > 0$. If ε is sufficiently small so that $d_1\lambda_0 < a - \varepsilon$, Proposition 2.1 enables us to deduce that

$$\liminf_{t\to\infty} u(\cdot, t) \geqq u^*(a-\varepsilon) \quad \text{uniformly in } \Omega. \tag{4.7}$$

Letting $\varepsilon \downarrow 0$ in (4.7) and using (4.5), we have

$$\lim_{t\to\infty} u(\cdot, t) = u^*(a) \quad \text{uniformly in } \Omega. \tag{4.8}$$

Thus (4.6) and (4.8) accomplish the proof. □

When (A.5) is assumed in place of (A.6), it is possible to show the global attractivity of $(0, v^*(b))$ along the same line as Theorem 4.2.

THEOREM 4.3. *In addition to* (A.1)-(A.4), *assume* (A.5). *If* $b \geqq b^*$ *and* $\bar{a}(b) > a$, *then every nonnegative solution of* (1.1), (1.2) *with* $v(\cdot, 0) \not\equiv 0$ *satisfies*

$$\lim_{t\to\infty} (u(\cdot, t), v(\cdot, t)) = (0, v^*(b)) \quad \textit{uniformly in } \bar{\Omega}.$$

Remark 4.1. We summarize our stability results of §§ 3 and 4 in the (a, b) parameter space (see Fig. 1). Assume that (A.1)-(A.6) are imposed on f and g. (Observe that f and g defined by (1.3) satisfy (A.1)-(A.6).) Theorems 4.1-4.3 ensure that $(0, 0)$, $(u^*(a), 0)$, and $(0, v^*(b))$ become global attractors for (1.1), (1.2) when (a, b) lies in regions I, II, and III, respectively. Therefore, there are no positive steady states for $(a, b) \in \text{I} \cup \text{II} \cup \text{III}$. This fact agrees with the result of Li [16], who has discussed prey-predator systems by setting almost the same assumptions as (A.1)-(A.6). Finally, we should say that, by Theorems 3.2, 3.3, and 3.5, trivial and semitrivial steady states are unstable if $(a, b) \in \text{IV}$.

5. Analysis of the stationary problem by bifurcation theory. In the subsequent sections we study positive steady states for (1.1), (1.2). There are several results on their existence; [1]-[3], [5], [8], [9], and [16]. Among them, Li [16, Thms. 1, 2] (cf.

Dancer [8], [9]) has established necessary and sufficient conditions for prey-predator diffusion systems (similar to ours) by using the degree theory. According to his work, it will be shown that, under assumptions (A.1)-(A.6), stationary problem (3.7)-(3.10) has a positive solution if and only if all the trivial and semitrivial solutions are unstable; that is, (a, b) satisfies $a > \bar{a}(b)$ and $b > \bar{b}(a)$. However, the stability of positive steady states is still an open problem. We will discuss their stability with use of the bifurcation theory, although most of the existence parts are already known.

It is now convenient to restate our preceding results from the standpoint of bifurcation theory (see Fig. 1). When a (respectively, b) is regarded as a bifurcation parameter, $(u^*(a), 0)$ (respectively, $(0, v^*(b))$) appears as a primary bifurcation from $(0, 0)$ at $a = a^*$ (respectively $b = b^*$) and the stability changes there. Moreover, $(u^*(a), 0)$ and $(0, v^*(b))$ lose their stability, when (a, b) crosses the C_1-curve and the C_2-curve, respectively. Therefore, positive solutions of (3.7)-(3.9) will be realized as secondary bifurcations from $(u^*(a), 0)$ or $(0, v^*(b))$. See the works of Blat and Brown [1]-[3] and Conway, Gardner, and Smoller [4], [5], where some bifurcation techniques are used.

We first discuss the secondary bifurcation from $(u^*(a), 0)$. Let $a > a^*$ be fixed and regard b as a parameter. Define a nonlinear operator $\mathbf{F}: \mathbf{Y} \times R \to \mathbf{X}$ by

$$\mathbf{F}(U; b) = \begin{pmatrix} A_1 u - au - uf(u, v) \\ A_2 v - bv - vg(u, v) \end{pmatrix} \quad \text{for } U = \begin{pmatrix} u \\ v \end{pmatrix} \in \mathbf{Y}. \tag{5.1}$$

Clearly, $\mathbf{F}(U^*; b) = 0$, where $U^* = {}^t(u^*(a), 0)$. As in the proof of Lemma 3.4, we employ the results of Crandall and Rabinowitz [7] to show bifurcation at $b = \bar{b}(a)$. In what follows, we sometimes write $\bar{b}$ in place of $\bar{b}(a)$. The Fréchet derivative of $\mathbf{F}(U; b)$ with respect to U at $(U, b) = (U^*, \bar{b})$ is given by

$$\mathbf{F}_U(U^*; \bar{b}) = \mathbf{L}_1(a, \bar{b}), \tag{5.2}$$

where $\mathbf{L}_1(a, b)$ is defined by (3.16). We will verify the assumptions of Lemma 1.1 of [7].

LEMMA 5.1. (i) $\dim N(\mathbf{F}_U(U^*; \bar{b}(a))) = 1$ *and* $N(\mathbf{F}_U(U^*; \bar{b}(a))) = \{{}^t(\varphi_1, \varphi_2)\}$ *with* $\varphi_1 < 0$, $\varphi_2 > 0$ *in* Ω *and* $\|\varphi_2\|_2 = 1$.

(ii) $\operatorname{codim} R(\mathbf{F}_u(U^*; \bar{b}(a))) = 1$. *Moreover,* ${}^t(h_1, h_2) \in R(\mathbf{F}_U(U^*; \bar{b}(a)))$ *if and only if* $(h_2, \varphi_2)_2 = 0$.

(iii)

$$\mathbf{F}_{Ub}(U^*; \bar{b}(a)) \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix} \notin R(\mathbf{F}_U(U^*; \bar{b}(a))).$$

Proof. In view of (5.2), we examine $N(\mathbf{L}_1(a, \bar{b}))$ and $R(\mathbf{L}_1(a, \bar{b}))$. Recall that zero is the least eigenvalue of $L_2(a, \bar{b})$ (see (3.17)), so that $\dim N(\mathbf{L}_1(a, \bar{b})) = 1$ because of the invertibility of $L_1(a)$ by Lemma 2.3. Let ${}^t(\varphi_1, \varphi_2) \in N(\mathbf{L}_1(a, \bar{b}))$. Then

$$L_1(a)\varphi_1 = u^*(a) f_v(u^*(a), 0)\varphi_2, \qquad L_2(a, \bar{b})\varphi_2 = 0.$$

Since φ_2 does not change sign, we may take φ_2 such that $\varphi_2 > 0$ in Ω and $\|\varphi_2\|_2 = 1$. By (A.1),

$$u^*(a) f_v(u^*(a), 0)\varphi_2 < 0 \quad \text{in } \Omega,$$

which, together with Lemma 2.3, implies

$$\varphi_1 = L_1(a)^{-1}(u^*(a) f_v(u^*(a), 0)\varphi_2) < 0 \quad \text{in } \Omega.$$

Thus we have shown (i).

If ${}^t(h_1, h_2)$ is in $R(\mathbf{L}_1(a, \bar{b}))$, there must be a solution (u, v) of

$$L_1(a)u - u^*(a) f_v(u^*(a), 0)v = h_1, \qquad L_2(a, \bar{b})v = h_2.$$

It is well known that the second equation has a solution v if and only if $(h_2, \varphi_2) = 0$. For such a solution v, the first equation has a unique solution u because of the invertibility of $L_1(a)$. Thus (ii) has been proved.

Finally, we observe that

$$\mathbf{F}_{Ub}(U^*; \bar{b})\begin{pmatrix}\varphi_1\\ \varphi_2\end{pmatrix} = \begin{pmatrix}0\\ -\varphi_2\end{pmatrix}.$$

Hence the assertion of (iii) easily follows from (ii). □

We are ready to prove the following bifurcation result.

THEOREM 5.2. *In addition to* (A.1)-(A.4), *assume* (A.6). *Let* $a > a^*$ *be fixed. Then there exists a positive number* δ *such that, for every* $b \in (\bar{b}(a), \bar{b}(a) + \delta)$, *there is a solution* $(\underline{u}, \underline{v})$ *of* (3.7)-(3.10) *with the following properties*:

(a) $u^*(a) > \underline{u} > 0$ *and* $\underline{v} > 0$ *in* Ω.

(b) $(\underline{u}, \underline{v})$ *is the only nontrivial solution of* (3.7)-(3.10) *near* $(u^*(a), 0)$.

(c) $(\underline{u}, \underline{v})$ *is asymptotically stable.*

Proof. We will prove this theorem along the same line as Conway, Gardner, and Smoller [5, Thm. 3.5] by applying the bifurcation theory of Crandall and Rabinowitz [7].

Because of Lemma 5.1, all the assumptions of Lemma 1.1 in [7] are satisfied, so that there exist a positive number s_0 and continuously differentiable functions ${}^t(w_1, w_2): (-s_0, s_0) \to \mathbf{Y}$ and $\beta: (-s_0, s_0) \to R$ with the following properties:

(i) $\beta(0) = 0$.

(ii) ${}^t(w_1(s), w_2(s)) \in R(\mathbf{L}_1(a, \bar{b}))$ and $w_1(0) = w_2(0) = 0$.

(iii) If

$$\begin{aligned} &\underline{u}(s) = u^*(a) + s(\varphi_1 + w_1(s)), \ \underline{v}(s) = s(\varphi_2 + w_2(s)),\\ &b(s) = \bar{b} + \beta(s), \end{aligned} \tag{5.3}$$

where φ_1, φ_2 are given in Lemma 5.1, then $\mathbf{F}(\underline{U}(s); b(s)) = 0$ with $\underline{U}(s) = {}^t(\underline{u}(s), \underline{v}(s))$; and

(iv) $(\underline{U}(s), b(s))$ is the unique nontrivial solution of $\mathbf{F}(U; b) = 0$ in a neighborhood of $(U^*, \bar{b})$.

This fact implies the existence of a nontrivial solution of (3.7)-(3.9) when (a, b) lies near the C_1-curve defined by $b = \bar{b}(a)$. Since $\varphi_1 < 0$ and $\varphi_2 > 0$ in Ω by Lemma 5.1, it follows from (5.3) that

$$u^*(a) > \underline{u}(s) > 0 \quad \text{and} \quad \underline{v}(s) > 0 \quad \text{in } \Omega$$

for sufficiently small $s > 0$.

As a next step, we will discuss the linearized stability of $(\underline{u}(s), \underline{v}(s))$. Let $\pi(s)$ denote the principal eigenvalue of $\mathbf{F}_U(\underline{U}(s); b(s))$. According to Theorem 1.16 of [7],

$$\lim_{s \to 0} \{-sb'(s)(\partial \nu_2 / \partial b)(a, \bar{b}) / \pi(s)\} = 1,$$

where $\nu_2(a, b)$ is defined by (3.17). Therefore,

$$\pi(s) = s\beta'(s)(1 + o(1)) \quad \text{for sufficiently small } s > 0. \tag{5.4}$$

We will show

$$\beta'(0) > 0 \tag{5.5}$$

to see the asymptotic stability of $(\underline{u}(s), \underline{v}(s))$ for sufficiently small $s > 0$. (Moreover, (5.5) tells us the direction of the bifurcation; $b(s) > \bar{b}$ for small $s > 0$.) Now

$\mathbf{F}(\underline{U}(s); b(s)) = 0$ implies that

$$L_2(a, \bar{b})\underline{v}(s) - \beta(s)\underline{v}(s) - \underline{v}(s)\{g(\underline{u}(s), \underline{v}(s)) - g(u^*(a), 0)\} = 0. \tag{5.6}$$

Since $\varphi_2 \in N(L_2(a, \bar{b}))$ and $(\underline{v}(s), \varphi_2)_2 = s$ by Lemma 5.1, taking the $L^2(\Omega)$-inner product of (5.4) with φ_2 leads to

$$\beta(s) = -((\varphi_2 + w_2(s))\{g(\underline{u}(s), \underline{v}(s)) - g(u^*(a), 0)\}, \varphi_2)_2. \tag{5.7}$$

Differentiation of (5.7) with respect to s gives

$$\beta'(0) = -\int_\Omega \{g_u(u^*(a), 0)\varphi_1 + g_v(u^*(a), 0)\varphi_2\}\varphi_2^2\, dx. \tag{5.8}$$

Thus (5.5) follows from (5.8) because $\varphi_1 < 0$ and $\varphi_2 > 0$ by Lemma 5.1, $g_u(u^*(a), 0) > 0$ by (A.2), and $g_v(u^*(a), 0) \leqq 0$ by (A.6).

The assertions of this theorem will be derived from the above results by regarding b as a bifurcation parameter rather than s (use (5.5)). □

The preceding argument is valid for studying the secondary bifurcation from $(0, v^*(b))$ at C_2-curve defined by $a = \bar{a}(b)$. Let $b > b^*$ be fixed and regard a as a bifurcation parameter. We define a mapping $\mathbf{G}: \mathbf{Y} \times R \to \mathbf{X}$ by

$$\mathbf{G}(V; a) = \begin{pmatrix} A_1 u - au - uf(u, v) \\ A_2 v - bv - vg(u, v) \end{pmatrix} \quad \text{for } V = \begin{pmatrix} u \\ v \end{pmatrix} \in \mathbf{Y}.$$

Clearly, $\mathbf{G}(V^*; a) = 0$ for all $a \geqq 0$ with $\mathbf{V}^* = {}^t(0, v^*(b))$. Correspondingly to Lemma 5.1, it is possible to show the following lemma.

LEMMA 5.3. (i) $\dim N(\mathbf{G}_V(V^*; \bar{a}(b))) = 1$ *and* $N(\mathbf{G}_V(V^*; \bar{a}(b))) = \{{}^t(\psi_1, \psi_2)\}$ *with* $\psi_1 > 0$ *and* $\psi_2 > 0$ *in* Ω *and* $\|\psi_1\|_2 = 1$.

(ii) $\operatorname{codim} R(\mathbf{G}_V(V^*; \bar{a}(b))) = 1$. *Moreover,* ${}^t(h_1, h_2) \in R(\mathbf{G}_v(V^*; \bar{a}(b)))$ *if and only if* $(h_1, \psi_1)_2 = 0$.

(iii) $$\mathbf{G}_{Va}(V^*; \bar{a}(b))\begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} \notin R(\mathbf{G}_V(V^*; \bar{a}(b))).$$

The following theorem can be shown in the same manner as Theorem 5.2 with the use of Lemma 5.3.

THEOREM 5.4. *In addition to* (A.1)-(A.4), *assume* (A.5). *Let* $b > b^*$ *be fixed. Then there exists a positive number* δ *such that, for every* $a \in (\bar{a}(b), \bar{a}(b) + \delta)$, *there exists a solution* $(\bar{u}, \bar{v})$ *of* (3.7)-(3.10) *that satisfies*:

(a) $\bar{u} > 0$ *and* $\bar{v} > v^*(b)$ *in* Ω.

(b) $(\bar{u}, \bar{v})$ *is the only nontrivial solution of* (3.7)-(3.10) *near* $(0, v^*(b))$.

(c) $(\bar{u}, \bar{v})$ *is asymptotically stable.*

Sketch of proof. Owing to Lemma 5.3, the bifurcation theory of Crandall and Rabinowitz yields the existence of a positive number s_0 and continuously differentiable functions ${}^t(w_1, w_2)$: $(-s_0, s_0) \to \mathbf{Y}$ and α: $(-s_0, s_0) \to R$ satisfying (i) $\alpha(0) = 0$; (ii) ${}^t(w_1(s), w_2(s)) \in R(\mathbf{G}_V(V^*; \bar{a}))$ and $w_1(0) = w_2(0) = 0$; and (iii) if

$$\bar{u}(s) = s(\psi_1 + w_1(s)), \qquad \bar{v}(s) = v^*(b) + s(\psi_2 + w_2(s)),$$
$$a(s) = \bar{a} + \alpha(s),$$

where ψ_1 and ψ_2 are as given in Lemma 5.3, then $\mathbf{G}(\bar{V}(s); a(s)) = 0$ with $\bar{V}(s) = {}^t(\bar{u}(s), \bar{v}(s))$. Therefore, in view of (iii), it is easy to show (a) because of the positivity of ψ_1 and ψ_2. In studying the asymptotic stability of $(\bar{u}(s), \bar{v}(s))$, it is essential to know the sign of $\alpha'(0)$. After some calculations, we can show that

$$\alpha'(0) = -\int_\Omega (f_u(0, v^*(b))\psi_1 + f_v(0, v^*(b))\psi_2)\psi_1^2\, dx \tag{5.9}$$

(cf. (5.8)). Hence (A.1), (A.5), and Lemma 5.3 enable us to see $\alpha'(0)>0$, which completes the proof. □

Theorems 5.2 and 5.4 assure the existence, uniqueness, and stability of positive solutions when $(a, b)\in$ IV lies in a neighborhood of C_1 or C_2 curve (see Fig. 1). To study the general case, we follow the argument of Ito [11, Thm. 5.1] based on the global bifurcation result of Rabinowitz [18] (see [1, Thm. 3.3]). Then it is possible to show Theorem 5.5.

THEOREM 5.5. *Assume* (A.1)-(A.6) *and let* $a>a^*$ *be fixed. Then there exists a branch of positive solutions of* (3.7)-(3.9) *that bifurcates from* $(u^*(a), 0)$ *at* $b=\bar{b}(a)$ *and meets with* $(0, v^*(b))$ *at* $b=(\bar{a})^{-1}(a)$, *where* $(\bar{a})^{-1}$ *is the inverse function of* $\bar{a}$.

Thus we have obtained a fairly clear understanding of the set of solutions for (3.7)-(3.10) and their stability properties. For example, consider the case $a>a^*$. Theorem 4.2 implies that $(u^*(a), 0)$ is a global attractor for (1.1), (1.2) if $b<\bar{b}(a)$. As b becomes larger than $\bar{b}(a)$, $(u^*(a), 0)$ loses its stability, and a positive solution, which is stable when b is near $\bar{b}(a)$, bifurcates from $(u^*(a), 0)$ at $b=\bar{b}(a)$ (Theorem 5.2). Such a positive solution exists for $\bar{b}(a)<b<(\bar{a})^{-1}(a)$ (Theorem 5.5), is stable when b is near $(\bar{a})^{-1}(a)$, and becomes identical with $(0, v^*(b))$ at $b=(\bar{a})^{-1}(a)$ (Theorem 5.4). By Theorem 4.3, $(0, v^*(b))$ is a global attractor for (1.1), (1.2) whenever b is larger than $(\bar{a})^{-1}(a)$.

6. Bifurcation from a double eigenvalue. In this section we discuss stability properties of positive solutions for (3.7)-(3.9) in the case when $(a, b)\in$ VI lies in a neighborhood of (a^*, b^*) $(\equiv(d_1\lambda_0, d_2\lambda_0))$.

The operator $\mathbf{L}_0(a, b)$ defined by (3.14) has zero as a double eigenvalue for $(a, b)=(a^*, b^*)$. We will derive appropriate expressions of positive solutions by simultaneously regarding a, b as bifurcation parameters. Throughout this section we assume, in addition to (A.1)-(A.6), that

$$f_u(0,0)<0 \quad \text{and} \quad g_v(0,0)<0. \tag{A.7}$$

Define a nonlinear mapping $\mathbf{H}:\mathbf{Y}\times R\to\mathbf{X}$ by

$$\mathbf{H}(U; a, b)=\begin{pmatrix} A_1u-au-uf(u,v)\\ A_2v-bv-vg(u,v)\end{pmatrix} \quad \text{for } U=\begin{pmatrix}u\\v\end{pmatrix}\in\mathbf{Y}. \tag{6.1}$$

Then $\mathbf{H}(0; a, b)=0$ for all a, b and $\mathbf{H}_U(0; a^*, b^*)=\mathbf{L}_0(a^*, b^*)$. In what follows, we simply write $\mathbf{L}_0$ in place of $\mathbf{L}_0(a^*, b^*)$. Clearly, dim $N(\mathbf{L}_0)=$ codim $R(\mathbf{L}_0)=2$. We can take $\{\Phi_1, \Phi_2\}\in N(\mathbf{L}_0)$ with

$$\Phi_1=\begin{pmatrix}\varphi_0\\0\end{pmatrix} \quad \text{and} \quad \Phi_2=\begin{pmatrix}0\\\varphi_0\end{pmatrix}, \tag{6.2}$$

where φ_0 is the eigenfunction of (3.11) corresponding to λ_0 and satisfies $\varphi_0>0$ in Ω and $\|\varphi_0\|_2=1$. Moreover,

$$\begin{pmatrix}h_1\\h_2\end{pmatrix}\in R(\mathbf{L}_0) \quad \text{if and only if } (h_1, \varphi_0)_2=(h_2, \varphi_0)_2=0. \tag{6.3}$$

For $U={}^t(u, v)\in\mathbf{X}$, we define

$$\mathbf{P}U=(u, \varphi_0)_2\Phi_1+(v, \varphi_0)_2\Phi_2$$

and decompose $\mathbf{X}$ as $\mathbf{X}=\mathbf{X}_1+\mathbf{X}_2$ with $\mathbf{X}_1=\mathbf{P}\mathbf{X}$ and $\mathbf{X}_2=(\mathbf{I}-\mathbf{P})\mathbf{X}$. Similarly, $\mathbf{Y}$ is decomposed as $\mathbf{Y}=\mathbf{Y}_1+\mathbf{Y}_2$ with $\mathbf{Y}_1=\mathbf{P}\mathbf{Y}$ and $\mathbf{Y}_2=(\mathbf{I}-\mathbf{P})\mathbf{Y}$. Then $\mathbf{X}_1=\mathbf{Y}_1=N(\mathbf{L}_0)$, $\mathbf{X}_2=R(\mathbf{L}_0)$, and $\mathbf{Y}_2=R(\mathbf{L}_0)\cap\mathbf{Y}$.

We will look for solutions of $\mathbf{H}(U; a, b) = 0$ in the form

$$U = s\{\cos\omega\Phi_1 + \sin\omega\,\Phi_2 + W\} \quad \text{with } w = {}^t(w_1, w_2) \in \mathbf{Y}_2, \tag{6.4}$$

where s and ω are parameters. Since positive solutions are concerned, we restrict ω to $(0, \pi/2)$. Let $\omega \in (0, \pi/2)$ be fixed for the time being. We define a nonlinear mapping $\mathbf{K}(W, \alpha, \beta; s): \mathbf{Y}_2 \times R \times R \times R \to \mathbf{X}$ by

$$\begin{aligned}\mathbf{K}(W, \alpha, \beta; s) &= s^{-1}\mathbf{H}(s(\cos\omega\Phi_1 + \sin\omega\Phi_2 + W); a^* + \alpha, b^* + \beta) \\ &= \mathbf{L}_0 W - \begin{pmatrix}\alpha(\varphi_0\cos\omega + w_1)\\ \beta(\varphi_0\sin\omega + w_2)\end{pmatrix} \\ &\quad - \begin{pmatrix}(\varphi_0\cos\omega + w_1)f(\tilde{u}, \tilde{v})\\ (\varphi_0\sin\omega + w_2)g(\tilde{u}, \tilde{v})\end{pmatrix}\end{aligned} \tag{6.5}$$

for $W = {}^t(w_1, w_2)$ with $\tilde{u} = s(\varphi_0\cos\omega + w_1)$ and $\tilde{v} = s(\varphi_0\sin\omega + w_2)$. Clearly, $\mathbf{K}$ is a C^1-mapping from $\mathbf{Y}_2 \times R \times R \times R$ to $\mathbf{X}$ satisfying $\mathbf{K}(0, 0, 0; 0) = 0$. The Fréchet derivative of $\mathbf{K}$ with respect to (W, α, β) at $(W, \alpha, \beta, s) = (0, 0, 0, 0)$ is the linear mapping

$$(\hat{W}, \hat{\alpha}, \hat{\beta}) \to \mathbf{L}_0\hat{W} - (\hat{\alpha}\cos\omega)\Phi_1 - (\hat{\beta}\sin\omega)\Phi_2,$$

which is an isomorphism from $\mathbf{Y}_2 \times R \times R$ to $\mathbf{X}$. Therefore, the Implicit Function Theorem implies the existence of continuously differentiable functions $(\tilde{W}(s), \tilde{\alpha}(s), \tilde{\beta}(s))$, defined for sufficiently small $|s|$, that satisfy (i) $\tilde{W}(0) = 0$, $\tilde{\alpha}(0) = \tilde{\beta}(0) = 0$ and (ii) $\mathbf{K}(\tilde{W}(s), \tilde{\alpha}(s), \tilde{\beta}(s); s) = 0$. Hence, if we set

$$\begin{aligned}&u(s) = s(\varphi_0\cos\omega + \tilde{w}_1(s), \qquad \tilde{v}(s) = s(\varphi_0\sin\omega + \tilde{w}_2(s)), \\ &\tilde{a}(s) = a^* + \tilde{\alpha}(s), \qquad \tilde{b}(s) = b^* + \tilde{\beta}(s),\end{aligned} \tag{6.6}$$

then $(\tilde{U}(s), \tilde{a}(s), \tilde{b}(s))$ with $\tilde{U}(s) = {}^t(\tilde{u}(s), \tilde{v}(s))$ becomes a nontrivial solution of $\mathbf{H}(U; a, b) = 0$. Especially, since $(\tilde{w}_i(s), \varphi_0)_2 = 0$ for $i = 1, 2$, it follows from $\mathbf{H}(\tilde{U}(s); \tilde{a}(s), \tilde{b}(s)) = 0$ that

$$\tilde{\alpha}(s)\cos\omega = -((\varphi_0\cos\omega + \tilde{w}_1(s))f(\tilde{u}(s), \tilde{v}(s)), \varphi_0)_2, \tag{6.7}$$

$$\tilde{\beta}(s)\sin\omega = -((\varphi_0\sin\omega + \tilde{w}_2(s))g(\tilde{u}(s), \tilde{v}(s)), \varphi_0)_2. \tag{6.8}$$

Differentiating (6.7) and (6.8) with respect to s, after some calculations we can derive

$$\tilde{\alpha}'(0) = -(f_u(0, 0)\cos\omega + f_v(0, 0)\sin\omega)\|\varphi_0\|_3^3, \tag{6.9}$$

$$\tilde{\beta}'(0) = -(g_u(0, 0)\cos\omega + g_v(0, 0)\sin\omega)\|\varphi_0\|_3^3, \tag{6.10}$$

so that

$$\begin{aligned}\lim_{s\to 0}\frac{\tilde{b}(s) - b^*}{\tilde{a}(s) - a^*} &= \lim_{s\to 0}\frac{\tilde{\beta}(s)}{\tilde{\alpha}(s)} \\ &= \frac{g_u(0, 0)\cos\omega + g_v(0, 0)\sin\omega}{f_u(0, 0)\cos\omega + f_v(0, 0)\sin\omega}.\end{aligned} \tag{6.11}$$

If the right-hand side of (6.11) is denoted by $l(\omega)$, then $l(\omega)$ is an increasing function of $\omega \in (0, \pi/2)$ such that

$$\lim_{\omega\to 0} l(\omega) = \frac{g_u(0, 0)}{f_u(0, 0)} \quad \text{and} \quad \lim_{\omega\to\pi/2} l(\omega) = \frac{g_v(0, 0)}{f_v(0, 0)}. \tag{6.12}$$

Now we recall some properties of C_1- and C_2-curves near (a^*, b^*). The C_1 curve, defined by $b = \bar{b}(a)$, satisfies

$$\lim_{a\to a^*}\frac{d\bar{b}}{da}(a) = \frac{g_u(0, 0)}{f_u(0, 0)} \tag{6.13}$$

by Lemma 3.4(iii), and the C_2-curve, defined by $a=\bar{a}(b)$, satisfies

$$\lim_{a\to a^*}\frac{d(\bar{a})^{-1}}{da}(a)=\frac{g_v(0,0)}{f_v(0,0)} \tag{6.14}$$

by Lemma 3.6(iii). In view of (6.12)-(6.14), we can conclude that, for each $\omega\in(0,\pi/2)$ sufficiently close to zero, the solution $(\tilde{u}(s),\tilde{v}(s))$ constructed as above coincides with the positive solution bifurcating from $(u^*(a),0)$ at the C_1-curve (see Theorem 5.2); and that, for each $\omega\in(0,\pi/2)$ sufficiently close to $\pi/2$, $(u(s),v(s))$ coincides with the positive solution bifurcating from $(0,v^*(b))$ at the C_2-curve (see Theorem 5.4). Thus we may see that the expression (6.6), for sufficiently small $s>0$, represents a branch of positive solutions of (3.7)-(3.10) that connects the bifurcating positive solution from the C_1-curve with that from the C_2-curve.

We will discuss the asymptotic stability of $(\tilde{u}(s),\tilde{v}(s))$ by the spectral analysis for $\mathbf{H}_U(\tilde{U}(s);\tilde{a}(s),\tilde{b}(s))$ along the idea of [6], [7]. Observe that the principal eigenvalue of $\mathbf{H}_U(\tilde{U}(0);\tilde{a}(0),\tilde{b}(0))=\mathbf{L}_0$ is zero with multiplicity 2. Hence, it suffices to examine the behavior near $s=0$ for (possibly two) eigenvalues $\zeta(s)$ such that

$$\mathbf{H}_U(\tilde{U}(s);\tilde{a}(s),\tilde{b}(s))U=\zeta(s)U \tag{6.15}$$

with $\zeta(0)=0$. We will look for eigenfunctions U in the form

$$U=\Phi_1+p\Phi_2+W,\qquad W\in Y_2, \tag{6.16}$$

where p and W are to be determined later. Substitution of (6.16) into (6.15) yields the following equivalent problem:

$$\mathbf{L}_0W-\mathbf{M}(s)W-(I-\mathbf{P})\mathbf{N}(s)(\Phi_1+p\Phi_2+W)-\zeta W=0, \tag{6.17}$$

$$-\tilde{\alpha}(s)-(\mathbf{N}(s)(\Phi_1+p\Phi_2+W),\Phi_1)_2-\zeta=0, \tag{6.18}$$

$$-\tilde{\beta}(s)p-(\mathbf{N}(s)(\Phi_1+p\Phi_2+W),\Phi_2)_2-\zeta p=0, \tag{6.19}$$

where

$$\mathbf{M}(s)=\begin{pmatrix}\alpha(s) & 0\\ 0 & \tilde{\beta}(s)\end{pmatrix},$$

$$\mathbf{N}(s)=\begin{pmatrix}(uf)_u(\tilde{u}(s),\tilde{v}(s)) & (uf)_v(\tilde{u}(s),\tilde{v}(s))\\ (vg)_u(\tilde{u}(s)),\tilde{v}(s)) & (vg)_v(\tilde{u}(s),\tilde{v}(s))\end{pmatrix}.$$

By (6.9) and (6.10),

$$\begin{aligned}\tilde{\alpha}(s)&=-s(f_u(0,0)\cos\omega+f_v(0,0)\sin\omega)\|\varphi_0\|_3^3+o(s),\\ \tilde{\beta}(s)&=-s(g_u(0,0)\cos\omega+g_v(0,0)\sin\omega)\|\varphi_0\|_3^3+o(s),\end{aligned} \tag{6.20}$$

for sufficiently small $s\geqq 0$. Moreover, some calculations give

$$\begin{aligned}(\mathbf{N}(s)\Phi_1,\Phi_1)_2&=s(2f_u(0,0)\cos\omega+f_v(0,0)\sin\omega)\|\varphi_0\|_3^3+o(s),\\ (\mathbf{N}(s)\Phi_2,\Phi_1)_2&=sf_v(0,0)\cos\omega\|\varphi_0\|_3^3+o(s),\\ (\mathbf{N}(s)\Phi_1,\Phi_2)_2&=sg_u(0,0)\sin\omega\|\varphi_0\|_3^3+o(s),\\ (\mathbf{N}(s)\Phi_2,\Phi_2)_2&=s(g_u(0,0)\cos\omega+2g_v(0,0)\sin\omega)\|\varphi_0\|_3^3+o(s).\end{aligned} \tag{6.21}$$

We first solve (W,ζ) from (6.17) and (6.18) as functions of s and p. From the Implicit Function Theorem there exist continuously differentiable functions $(W(s;p),\zeta(s;p))$, defined for small $s\geqq 0$, with the following properties:

(i) $(W(0;p),\zeta(0;p))=(0,0)$;

(ii) $(W(s;p),\zeta(s;p))$ satisfies (6.17) and (6.18);

(iii) $\|W(s;p)\|_{Y_2}\leqq Cs(1+|p|)$ and $|\zeta(s;p)|\leqq Cs(1+|p|)$ with some $C>0$.

Therefore, it follows from (6.18), with use of (6.20) and (6.21), that

$$\begin{aligned}\zeta(s;p) = &-sp(f_v(0,0)\cos\omega\|\varphi_0\|_3^3 + o(1))\\ &-s(f_u(0,0)\cos\omega\|\varphi_0\|_3^3 + o(1)).\end{aligned} \tag{6.22}$$

Moreover, since $(W(s;p), \zeta(s;p))$ must satisfy (6.19), we are led to the following equation:

$$\begin{aligned}0 &= -\zeta(s;p)p - \tilde{\beta}(s)p - (\mathbf{N}(s)(\Phi_1 + p\Phi_2 + W(s;p)), \Phi_2)_2\\ &= sp^2\{f_v(0,0)\cos\omega\|\varphi_0\|_3^3 + o(1)\}\\ &\quad + sp\{f_u(0,0)\cos\omega\|\varphi_0\|_3^3 - g_v(0,0)\sin\omega\|\varphi_0\|_3^3 + o(1)\}\\ &\quad - s\{g_u(0,0)\sin\omega\|\varphi_0\|_3^3 + o(1)\}.\end{aligned}$$

Hence,

$$\begin{aligned}&p^2\{f_v(0,0)\cos\omega + o(1)\} + p\{f_u(0,0)\cos\omega - g_v(0,0)\sin\omega + o(1)\}\\ &\quad - \{g_u(0,0)\sin\omega + o(1)\} = 0,\end{aligned}$$

so that we can find two continuous functions $p_\pm(s)$ such that

$$\begin{aligned}p_\pm(s) = (2f_v(0,0)\cos\omega)^{-1}&[(g_v(0,0)\sin\omega - f_u(0,0)\cos\omega)\\ &\pm\{(f_u(0,0)\cos\omega - g_v(0,0)\sin\omega)^2\\ &\quad + 4f_v(0,0)g_u(0,0)\sin\omega\cos\omega\}^{1/2}] + o(1).\end{aligned} \tag{6.23}$$

Substitution of (6.23) into (6.22) gives two eigenvalues $\zeta_\pm(s)$ such that

$$\begin{aligned}\zeta_\pm(s) = -2^{-1}s&[f_u(0,0)\cos\omega + g_v(0,0)\sin\omega\\ &\pm\{(f_u(0,0)\cos\omega - g_v(0,0)\sin\omega)^2\\ &\quad + 4f_v(0,0)g_u(0,0)\sin\omega\cos\omega\}^{1/2}] + o(s).\end{aligned}$$

Since it is easy to see $\operatorname{Re}\zeta_\pm(s) > 0$ for sufficiently small $s > 0$, the spectrum of $\mathbf{H}_U(\tilde{U}(s); \tilde{a}(s), \tilde{b}(s))$ lies in the right half-plane of $\mathbf{C}$. Thus we have shown the following result.

THEOREM 6.1. *If $(a, b) \in$ IV lies in a neighborhood of (a^*, b^*), then there exists a solution of* (3.7)-(3.10), *which is positive and asymptotically stable.*

Remark 6.1. If we take f and g of the form (1.3), then all assumptions (A.1)-(A.7) are fulfilled. Therefore, our results can be summarized as in § 1 (see Fig. 1) and improve on those of Blat and Brown [1]. Especially, the stability properties of the trivial and semitrivial steady-state solutions are completely determined.

Remark 6.2. Our stability analysis developed in this paper is valid with slight modification even if some of (A.1)-(A.7) are replaced by other suitable conditions. For example, it is possible to study the stability of positive steady states for reaction-diffusion systems such as the Holling-Tanner model (see [2]) or the competition model.

REFERENCES

[1] J. BLAT AND K. J. BROWN, *Bifurcation of steady-state solutions in predator-prey and competition systems*, Proc. Roy. Soc. Edinburgh Sect. A, 97 (1984), pp. 21-34.

[2] ———, *Global bifurcation of positive solutions in some systems of elliptic equations*, SIAM J. Math. Anal., 17 (1986), pp. 1339-1353.

[3] K. J. BROWN, *Nontrivial solutions of predator-prey systems with small diffusion*, Nonlinear Anal., 11 (1987), pp. 685-689.

[4] E. CONWAY, *Diffusion and the predator-prey interaction; steady states with flux at the boundaries*, in Nonlinear Partial Differential Equations, J. A. Smoller, ed., Contemporary Mathematics 17, American Mathematical Society, Providence, RI, 1983.

[5] E. CONWAY, R. GARDNER, AND J. SMOLLER, *Stability and bifurcation of steady-state solutions for predator-prey equations*, Adv. in Appl. Math., 3 (1982), pp. 288-334.

[6] M. G. CRANDALL AND P. H. RABINOWITZ, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321-340.

[7] ———, *Bifurcation, perturbation of simple eigenvalues, and linearized stability*, Arch. Rational Mech. Anal., 52 (1973), pp. 161-180.

[8] E. N. DANCER, *On positive solutions of some pairs of differential equations*, Trans. Amer. Math. Soc., 284 (1984), pp. 729-743.

[9] ———, *On positive solutions of some pairs of differential equations*, II, J. Differential Equations, 60 (1985), pp. 236-258.

[10] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, New York, 1984.

[11] M. ITO, *Global aspect of steady-states for competitive-diffusive systems with homogeneous Dirichlet conditions*, Phys. D, 14 (1984), pp. 1-28.

[12] T. KATO, *Perturbation Theory for Linear Operators*, Second edition, Springer-Verlag, Berlin, New York, 1980.

[13] H. KIELHÖFER, *On the Lyapunov-stability of stationary solutions of semilinear parabolic differential equations*, J. Differential Equations, 22 (1976), pp. 193-208.

[14] P. KORMAN AND A. W. LEUNG, *A general monotone scheme for elliptic systems with applications to ecological models*, Proc. Roy. Soc. Edinburgh Sect. A, 102 (1986), pp. 315-325.

[15] A. LEUNG, *Monotone schemes for semilinear elliptic systems related to ecology*, Math. Meth. Appl. Sci., 4 (1982), pp. 272-285.

[16] L. LI, *Positive solutions of steady states of predator-prey systems*, Nonlinear Analysis and Applications, V. Lakshmikantham, ed., Lecture Notes in Pure and Appl. Math. 109, Marcel Dekker, New York, 1987.

[17] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Springer-Verlag, Berlin, New York, 1984.

[18] P. H. RABINOWITZ, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487-513.

[19] F. ROTHE, *Uniform bounds from bounded L_p-functionals in reaction-diffusion equations*, J. Differential Equations, 45 (1982), pp. 207-233.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 346-361, March 1990
© 1990 Society for Industrial and Applied Mathematics
004

PATTERN FORMATION IN HETEROGENEOUS REACTION-DIFFUSION-ADVECTION SYSTEMS WITH AN APPLICATION TO POPULATION DYNAMICS*

S.-I. EI† AND M. MIMURA†

Abstract. Heterogeneous reaction-diffusion-advection equations are proposed for studying pattern formation due to spatial heterogeneity. The equations contain a small parameter ε, indicating the ratio of the diffusion and advection rates and the reaction rate. Two-timing methods in the limit $\varepsilon \downarrow 0$ make it possible to reduce the original partial differential equation problem to the approximating ordinary differential equation problem, so that asymptotic states of solutions can be investigated. As an application to population dynamics, population models of the Lotka-Volterra type are considered for studying the effect of spatial heterogeneity on development of spatial and temporal distributions of individuals.

Key words. pattern formation due to heterogeneities, two-timing method

AMS(MOS) subject classifications. 34A05, 92A15

1. Introduction. To theoretically understand spatial and temporal distributions of ecological populations, there have been proposed a large number of mathematical models that essentially include two terms, such as dispersal and growth and/or death. As one of these models, we have the following reaction-diffusion-advection system:

$$\frac{\partial u}{\partial t} + \operatorname{div} J = f(u), \quad t > 0, \quad x \in \Omega, \tag{1.1}$$

where

$$u = (u_1, u_2, \cdots, u_n), \qquad f(u) = (f_1(u), f_2(u), \cdots, f_n(u)),$$
$$J = J(u) = -(d_1 \nabla u_1 + u_1 \nabla e_1, d_2 \nabla u_2 + u_2 \nabla e_2, \cdots, d_n \nabla u_n + u_n \nabla e_n),$$

and Ω is a bounded domain in $\mathbb{R}^N$. Here u_i is the population density of the ith species at time t and position x, with the diffusion rate d_i, the tactic advection ∇e_i, and the growth term f_i. We conveniently write $(d_1, d_2, \cdots, d_n)$ and $(e_1, e_2, \cdots, e_n)$ as d and e, respectively. Usually d, e, and f depend on t, x, and u. The flux J consists of two dispersal processes: the normal random movement of individuals, and the directed movement toward the favorable region (or, on the contrary, from the unfavorable one). If e is constant, namely, if the advection term is absent, J is reduced to the usual diffusive flux. The boundary condition to (1.1) is

$$\langle \nu, J \rangle = 0_n, \quad t > 0, \quad x \in \partial\Omega \tag{1.2}$$

where $\langle , \rangle$ is an inner product, ν is the outward normal vector on the boundary $\partial\Omega$, and 0_n is the n-dimensional zero vector. This condition ecologically implies that there is no population flow through the boundary $\partial\Omega$.

The scalar case of (1.1), (1.2) ($n = 1$) has been widely investigated when the advection term is absent and the growth term f depends only on u. The resulting problem of (1.1), (1.2) is

$$\frac{\partial u}{\partial t} = \operatorname{div}(d(x)\nabla u) + f(u), \qquad t > 0, \quad x \in \Omega \tag{1.3}$$

* Received by the editors September 6, 1988; accepted for publication (in revised form) March 13, 1989.
† Department of Mathematics, Hiroshima University, Hiroshima 730, Japan.

with the boundary condition

$$\frac{\partial u}{\partial \nu}=0, \quad t>0, \quad x\in\partial\Omega. \tag{1.4}$$

For (1.3), (1.4), there are interesting results on existence, stability and bifurcation properties of nonconstant stationary solutions [1], [10], [11], [14]-[16], [22], [33], [39]. The most noteworthy result is that when the domain Ω is convex, any stable nonconstant stationary solutions do not exist if d is constant, but possibly exist due to suitable heterogeneities of $d(x)$. In application points of view, a class of equations such as (1.1) has been discussed by various investigators in population genetics [7], [8], in population dynamics [12], [25], [26], [29] and in neurobiology [20], [30], [31].

To our knowledge, there has not yet been any full and systematic analysis for the system of equations of (1.1) ($n\geqq 2$) to study the effect of spatial heterogeneity on solutions, except for several works including [13], [37]. Under such circumstances, Shigesada [34] has dealt with the extreme case, where the rate of dispersal is sufficiently large compared with that of growth. This restriction is required by the ecological situation such that the dispersal processes take place daily, while the growth process takes place only once or twice a year. Under this assumption, (1.1) is written as

$$\frac{\partial u}{\partial t}+\operatorname{div} J=\varepsilon f, \quad t>0, \quad x\in\Omega \tag{1.1}_\varepsilon$$

with a small parameter ε. By using the two-time-scale (two-timing) method, approximating ordinary differential equations can be formally derived from $(1.1)_\varepsilon$, (1.2) as $\varepsilon\downarrow 0$ [34]. Although this reduction of partial differential equations (PDEs) to ordinary differential equations (ODEs) is formal, we can obtain much information on the time development of solutions of $(1.1)_\varepsilon$, (1.2) by studying the corresponding ODEs. Recently, Ei and Mimura [5] and Ei [4] have shown that this reduction is valid for all t up to $O(1/\varepsilon)$, and that in some situations it is valid for all $t\in[0,\infty)$. Such reduction methods for infinite-dimensional dynamical systems to finite-dimensional dynamical systems are among the most recent interesting topics in the analysis of dynamical systems (see [2] and [9], for instance).

Let us show one population model equation, described by $(1.1)_\varepsilon$, (1.2), which represents competition between two species in the one-dimensional heterogeneous habitat. It is given by

$$\begin{aligned}
&\frac{\partial u_1}{\partial t}=\frac{\partial}{\partial x}\left(d_1\frac{\partial u_1}{\partial x}+u_1\dot{e}_1\right)+\varepsilon\left(r_1-\frac{\alpha_1u_1+\beta_1u_2}{K_1}\right)u_1,\\
&\frac{\partial u_2}{\partial t}=\frac{\partial}{\partial x}\left(d_2\frac{\partial u_2}{\partial x}+u_2\dot{e}_2\right)+\varepsilon\left(r_2-\frac{\beta_2u_1+\alpha_2u_2}{K_2}\right)u_2,\\
&\qquad\qquad\qquad\qquad\qquad\qquad t>0, \qquad x\in\Omega\equiv(0,1)
\end{aligned} \tag{1.5}_\varepsilon$$

with

$$d_i\frac{\partial u_i}{\partial x}+u_i\dot{e}_i=0, \quad t>0, \quad x\in\partial\Omega \quad (i=1,2), \tag{1.6}$$

where $\dot{e}_i=(de_i/dx)$ $(i=1,2)$ and (r_1, r_2) is the intrinsic growth rate, (α_1, α_2) and (β_1, β_2) are, respectively, the intraspecific and interspecific competition rates between the two species, and (K_1, K_2) is the carrying capacity. Here we assume that all the parameters are positive constants, leaving e_i and K_i $(i=1,2)$ to be functions of x. We

specify that $K_i(x)$ $(i=1,2)$ be humped as shown in Fig. 1, where x_i is the maximum point of $K_i(x)$ $(i=1,2)$, and consider the effect of the heterogeneities of $K_i(x)$ on the stability of spatial distributions of the two competing species. To do so, we introduce one parameter $\theta = x_2 - x_1$ that indicates the distance between the most favorable regions of the two competing species. We assume that e_i satisfies

$$e_i(x) = -\gamma \log K_i(x)$$

with a positive constant γ $(i=1,2)$. This relation implies that each species tends to migrate toward the higher gradient of its carrying capacity in Ω. This requirement seems to be phenomenologically reasonable. We show numerically the time development of (u_1, u_2) of $(1.5)_\varepsilon$, (1.6) for the same initial distributions. When $\theta = 0$, u_1 exists while u_2 is extinct (Fig. 2(a)). When $\theta = 0.2$, u_2 exists while u_1 is extinct (Fig. 2(b)). On the other hand, when $\theta = 0.4$, the situation changes so that u_1 and u_2 can coexist (Fig. 2(c)). These simulations suggest that stable spatial distributions of the competing species strongly depend on the heterogeneous carrying capacities. Mathematically speaking, bifurcation phenomena may possibly occur when θ varies.

Motivated by these phenomena, we are interested in pattern formation of solutions of $(1.1)_\varepsilon$, (1.2). Especially, we study the dependency of spatial heterogeneities of J and f on solutions.

In § 2, we apply the two-timing method to the PDE problem $(1.1)_\varepsilon$, (1.2) and derive the approximating ODE problem ((2.8), (2.9)) in the limit $\varepsilon \downarrow 0$. The validity of this method is also shown. In § 3, we study existence and stability problems of stationary solutions as well as periodic solutions of $(1.1)_\varepsilon$, (1.2). We emphasize here that not only existence but also stability of such solutions of the PDE problem generically inherit from those of the approximating ODE problem. In § 4, we give proofs of the theorems shown in § 3. Finally, in § 5, as an application of our procedure, we study the qualitative behavior of solutions of the specified model $(1.5)_\varepsilon$, (1.6) when θ varies. Further application to population models with spatially heterogeneous environments will be reported in [6].

2. Reduction of PDE problems to the approximating ODE problems. In this section, we apply the two-timing method ([33], for instance) to the following initial boundary value problem to $(1.1)_\varepsilon$ with a small parameter ε:

$$\frac{\partial u}{\partial t} + \operatorname{div} J = \varepsilon f(x, u), \quad t > 0, \quad x \in \Omega, \tag{2.1}_\varepsilon$$

$$\langle \nu, J \rangle = 0_n, \quad t > 0, \quad x \in \partial\Omega, \tag{2.2}$$

$$u(0, x) = \xi(x), \qquad x \in \bar{\Omega}, \tag{2.3}$$

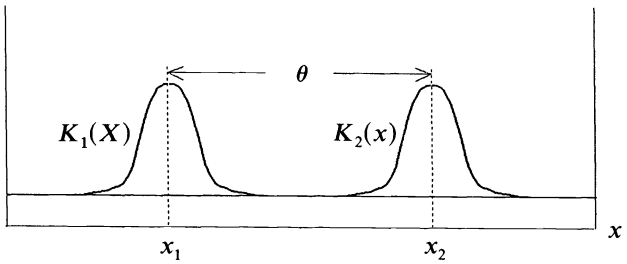


FIG. 1. *Graph of heterogeneous carrying capacity* K_i $(i=1,2)$.

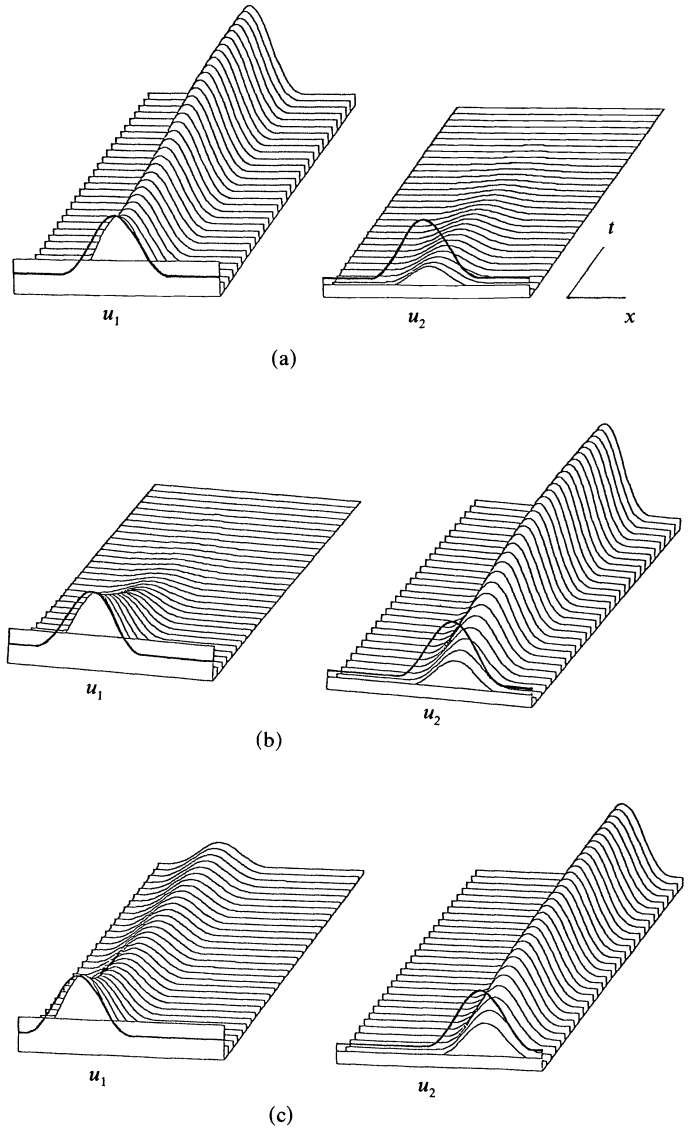


FIG. 2. *Time development of solutions of* $(1.5)_\varepsilon$, (1.6) *with the same initial condition for the suitably specified parameters (see* (4.5), (4.6) *and* $\varepsilon = 0.1$) *when* θ *varies*: (a) $\theta = 0$, (b) $\theta = 0.2$, (c) $\theta = 0.4$. *In each figure, bold lines denote the functional forms of* $K(x)$.

where $\xi = (\xi_1, \xi_2, \cdots, \xi_n)$, Ω is a bounded domain in $\mathbb{R}^N$ with sufficiently smooth boundary $\partial\Omega$, and e and f are smooth functions in every argument. Here we assume that

(A) d is a positive constant and e is a function of x only.

When ε is sufficiently small, namely, when the dispersal proceeds much faster than the dynamics, we can imagine that the spatial distributions of individuals are governed by the dispersal process as the initial stage and then are followed by the dynamical process as the second stage. Such time development of solutions requires us to introduce two different timescales t and τ $(=\varepsilon t)$. We now attempt to find solutions of $(2.1)_\varepsilon$, (2.2) in the form

$$u(t, x; \varepsilon) = u_0(t, \tau, x) + \varepsilon u_1(t, \tau, x) + O(\varepsilon^2). \tag{2.4}$$

With the relation $\partial/\partial t = \partial/\partial t + \varepsilon(\partial/\partial\tau)$, inserting (2.4) into $(2.1)_\varepsilon$-(2.3) and equating coefficients of like powers of ε, we obtain

$$\begin{aligned} &\frac{\partial}{\partial t} u_0 + \operatorname{div} J_0 = 0, \quad t>0, \quad x\in\Omega, \\ &\langle \nu, J_0\rangle = 0_n, \quad t>0, \quad x\in\partial\Omega, \end{aligned} \tag{2.5}$$

where $J_0 = J(u_0)$. Similarly,

$$\begin{aligned} &\frac{\partial}{\partial t} u_1 + \frac{\partial}{\partial \tau} u_0 + \operatorname{div} J_1 = f(x, u_0), \quad t>0, \quad x\in\Omega, \\ &\langle \nu, J_1\rangle = 0_n, \quad t>0, \quad x\in\partial\Omega, \end{aligned} \tag{2.6}$$

where $J_1 = J(u_1)$. Let $\varphi(x)$ be the stationary solution of (2.5) with $\int_\Omega \varphi(x)\,dx = 1_n$, where 1_n is the n-dimensional vector whose components are all 1. Then we find that there is a function $U(\tau) = (U_1, U_2, \cdots, U_n)(\tau)$ such that

$$\lim_{t\to\infty} u_0(t, \tau, x) = U(\tau)\cdot\varphi(x) \tag{2.7}$$

for a uniquely determined φ. Here, for $x = (x_1, x_2, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n)$, $x\cdot y$ means the vector $(x_1y_1, x_2y_2, \cdots, x_ny_n)$. Similarly, $V\cdot x$ means the set $\{v\cdot x \in \mathbb{R}^n \mid v\in V\}$ for $x\in\mathbb{R}^n$ and a set V in $\mathbb{R}^n$. Here $U(\tau)$ is determined as follows. Integrating (2.6) over Ω, we have

$$\frac{\partial}{\partial t}\int_\Omega u_1(t, \tau, x)\,dx + \frac{dU}{d\tau} = \int_\Omega f(x, u_0(t, \tau, x))\,dx. \tag{2.8}$$

Suppose that for any fixed τ, $u_1(t, \tau, x)$ is bounded for all $t>0$, or equivalently, $\partial/\partial t\int_\Omega u_1(t, \tau, x)\,dx \to 0$ as $t\to\infty$ (Ei [4]). Then, when $t\to\infty$, (2.8) reduces to the following ODE of U only:

$$\frac{dU}{d\tau} = F(U), \qquad \tau>0, \tag{$2.9)_1$}$$

where $F(U) = \int_\Omega f(x, U\cdot\varphi(x))\,dx = (F_1(U), F_2(U), \cdots, F_n(U))$. Since $\partial/\partial t\int_\Omega \cdot u_0(t, \tau, x)\,dx = 0_n$ by (2.5), we have $\int_\Omega u_0(0, \tau, x)\,dx = U(\tau)$. Therefore, by $u_0(0, 0, x) = \xi(x)$, the initial condition to $(2.9)_1$ is

$$U(0) = \int_\Omega \xi(x)\,dx. \tag{$2.9)_2$}$$

Thus the initial value problem for $U(\tau)$ is formulated as (2.9).

Remark 2.1. If (2.7) is valid, the spatial and temporal distributions of solutions of $(2.1)_\varepsilon$-(2.3) for large time t could be governed by $\varphi(x)$ and $U(\tau)$, respectively.

Remark 2.2. If $e_i(x)$ are all constant, that is, if $\nabla e_i(x)\equiv 0$, then $\varphi_i(x) = 1/|\Omega|$, so that $F_i(U) = 1/|\Omega|\int_\Omega f_i(x, U_1/|\Omega|, \cdots, U_n/|\Omega|)\,dx$. This case has already been discussed by Conway, Hoff, and Smoller [3], Yu [37], and Hale [13].

We omit the construction of $u_0(t, \tau, x)$, because it is shown in Ei [4]. However, we should note here that it is not unique. One explicit form is $u_0(t, \tau, x) = w(t, x) + U(\tau)\cdot\varphi(x)$, where $w(t, x)$ is the solution of

$$\begin{aligned} &\frac{\partial w}{\partial t} + \operatorname{div} J(w) = 0_n, \quad t>0, \quad x\in\Omega, \\ &w(0, x) = \xi(x) - \int_\Omega \xi(x)\,dx\cdot\varphi(x), \qquad x\in\bar{\Omega}. \end{aligned} \tag{2.10}$$

Next we briefly mention the validity of $u_0(t, \tau, x)$ as an approximation to the exact solution $u(t, x; \varepsilon)$ of $(2.1)_\varepsilon$-(2.3). We first introduce some notation. Let $C^0(\Omega; \mathbb{R}^n)$ be the Banach space consisting of all bounded continuous functions over Ω with the norm $\|u\|_0 = \sup_{x\in\Omega} |u(x)|$ for $u \in C^0(\Omega; \mathbb{R}^n)$ and $\operatorname{dist}\{V_1, V_2\} = \inf\{\|x-y\|_0 | x \in V_1, y \in V_2\}$ for two closed sets V_1, V_2 in $C^0(\Omega; \mathbb{R}^n)$. Similarly, $C^k(\Omega; \mathbb{R}^n)$ denotes the Banach space consisting of all k-times continuously differentiable functions over Ω with the norm $\|u\|_k = \sum_{i=0}^k \sup_{x\in\Omega} |D^i u(x)|$. When a finite-dimensional space can be considered as a subspace of $C^0(\Omega; \mathbb{R}^n)$, we identify the Euclidean norm of the finite-dimensional space with the restriction of $\|\cdot\|_0$ to the space, say $|\cdot|$, because all norms of a finite-dimensional space are equivalent to each other. Similarly, we use the same symbol $\operatorname{dist}\{V_1, V_2\}$ even if V_1 and V_2 are subsets of the finite-dimensional space.

Let $\pi(\tau; \Xi)$ be a solution of $(2.9)_1$ with the initial condition

$$U(0) = \Xi \in \mathbb{R}^n. \tag{2.11}$$

DEFINITION. We call a closed bounded set Γ in $\mathbb{R}^n$ an *exponentially stable attractor* of π if there exist an open bounded set $V (\supset \Gamma)$ in $\mathbb{R}^n$ and constants $M_0 > 0$, $\alpha > 0$ such that

$$\operatorname{dist}(\pi(\tau; \xi), \Gamma) \leqq M_0 e^{-\alpha\tau} \operatorname{dist}(\xi, \Gamma)$$

for any $\xi \in V$ and any $\tau > 0$.

THEOREM 2.1 [4]. *If $\pi(\tau; \int_\Omega \xi(x)\, dx)$ converges, as $\tau \to \infty$, to some exponentially stable attractor Γ, then there exist constants $\varepsilon_1 > 0$ and $C_1 > 0$ such that for any $\varepsilon \in (0, \varepsilon_1]$ a solution of $(2.1)_\varepsilon$-(2.3) $u(t, x; \varepsilon)$ exists for all t and*

$$\operatorname{dist}\{u(t, \cdot; \varepsilon), \Gamma \cdot \varphi(\cdot)\} \leqq C_1 \varepsilon$$

for sufficiently large t.

Theorem 2.1 indicates that if the initial value $\int_\Omega \xi(x)\, dx$ of $(2.9)_1$ is contained in the attractive region of Γ, then $u(t, \cdot; \varepsilon)$ eventually enters an ε-neighborhood of $\Gamma \cdot \varphi(\cdot)$ in $C^0(\Omega; \mathbb{R}^n)$.

As a special case, if Γ consists of either only one equilibrium or only one periodic orbit of $(2.9)_1$, Theorem 2.1 can be stated as follows.

COROLLARY 2.1 [4]. *In addition to the assumptions of Theorem 2.1, assume that Γ is an equilibrium of $(2.9)_1$. Then there exist constants $\varepsilon_1 > 0$ and $C_1 > 0$ such that*

$$\|u(t, \cdot; \varepsilon) - u_0(t, \varepsilon t, \cdot)\|_0 \leqq C_1 \varepsilon$$

uniformly for any $\varepsilon \in (0, \varepsilon_1]$ and any $t \in [0, \infty)$.

COROLLARY 2.2 [4]. *In addition to the assumptions of Theorem 2.1, assume that Γ is a periodic orbit of $(2.9)_1$. Then, there exist constants $\varepsilon_1 > 0$, $\eta \in (0, 1)$, and $C_1 > 0$ such that*

$$\operatorname{dist}\left\{u(t, \cdot; \varepsilon), \bigcup_{s\geqq 0} u_0(s, \varepsilon s, \cdot)\right\} \leqq C_1 \varepsilon^\eta$$

uniformly for any $\varepsilon \in (0, \varepsilon_1]$ and any $t \in [0, \infty)$.

The results above indicate that the solution $u(t, x; \varepsilon)$ of $(2.1)_\varepsilon$-(2.3) asymptotically enters an ε-neighborhood of the lowest approximation $u_0(t, \varepsilon t, x)$ in suitable function spaces. However, these do not reveal the behavior of the solution in that neighborhood. That is, even if Γ consists of a unique periodic orbit, for instance, we cannot say whether $(2.1)_\varepsilon$, (2.2) also has a unique periodic solution such that the solution $u(t, x; \varepsilon)$ tends to it as $t \to \infty$.

In the next section, we study the relation between the asymptotic states of solutions of (2.9) and $(2.1)_\varepsilon$, (2.2) when ε is sufficiently small.

3. Asymptotic states of $u(t, x; \varepsilon)$. In this section, to study asymptotic states of $u(t, \cdot; \varepsilon)$ of $(2.1)_\varepsilon$, (2.2), we are concerned with the existence and stability of stationary solutions as well as periodic solutions.

DEFINITION. Let π_e and $\pi_p = \{\pi_p(\tau)\}$ be an equilibrium and a periodic orbit of $(2.9)_1$, respectively. Consider the linearized equation of $(2.9)_1$ about π_e or π_p:

$$\frac{dy}{d\tau} = F_U y, \tag{3.1}$$

where $F_U = dF/dU$. We call π_e a *nondegenerate* equilibrium if $F_U(\pi_e)$ has no zero eigenvalue, and call π_p a *nondegenerate* periodic orbit if the characteristic multiplier 1 is an isolated simple eigenvalue of the period map of (3.1).

THEOREM 3.1. *If π_e is a nondegenerate equilibrium of $(2.9)_1$, there is a constant $\varepsilon_0 > 0$ such that $(2.1)_\varepsilon$, (2.2) has a unique stationary solution $\bar{u}(x; \varepsilon)$ for $|\varepsilon| < \varepsilon_0$ which satisfies $\bar{u}(\cdot; \varepsilon) \in C((-\varepsilon_0, \varepsilon_0); C^0(\Omega; \mathbb{R}^n))$ and $\lim_{\varepsilon \downarrow 0} \bar{u}(x; \varepsilon) = \pi_e \cdot \varphi(x)$ uniformly in $x \in \Omega$.*

THEOREM 3.2. *Suppose that π_e is a nondegenerate equilibrium of $(2.9)_1$ and that $\bar{u}(\cdot; \varepsilon)$ is a unique stationary solution associated with π_e as shown in Theorem 3.1. Let L_ε be a linearized operator of $(2.1)_\varepsilon$, (2.2) about $\bar{u}(\cdot; \varepsilon)$. Then there exist $\varepsilon_1 > 0$, $C_1 > 0$, and $\alpha > 0$ such that the spectrum of L_ε consists of two spectral sets $\sigma_1(\varepsilon)$, $\sigma_2(\varepsilon)$ satisfying $\sigma_1(\varepsilon) \subset \{\lambda \in \mathbb{C} \,|\, |\lambda| \leqq C_1 \varepsilon\}$ and $\sigma_2(\varepsilon) \subset \{\lambda \in \mathbb{C} \,|\, \operatorname{Re} \lambda < -\alpha\}$ for $0 < \varepsilon \leqq \varepsilon_1$. For any $\lambda(\varepsilon) \in \sigma_1(\varepsilon)$, there is some $\lambda_0 \in \sigma(F_U(\pi_e))$ such that $\lim_{\varepsilon \downarrow 0} \lambda(\varepsilon)/\varepsilon = \lambda_0$ and its converse also holds, where $\sigma(F_U(\pi_e))$ denotes the spectrum of $F_U(\pi_e)$.*

For this theorem, (i) L_ε has n eigenvalues in an ε-neighborhood of the origin; (ii) by the suitable ε-scaling, these eigenvalues coincide with those of $F_U(\pi_e)$.

THEOREM 3.3. *If the nondegenerate equilibrium π_e is asymptotically stable, $\bar{u}(x; \varepsilon)$ associated with π_e is also asymptotically stable for sufficiently small $\varepsilon > 0$. Moreover, it has an attractive region V, which is independent of ε, in $C^0(\Omega; \mathbb{R}^n)$.*

Similar results hold for the case of the nondegenerate periodic orbit.

THEOREM 3.4. *If π_p is a nondegenerate periodic orbit with period τ_0, there is $\varepsilon_0 > 0$ such that $(2.1)_\varepsilon$, (2.2) has a periodic solution $P(t, x; \varepsilon)$ with period t_ε for $0 < \varepsilon < \varepsilon_0$ which satisfies*

$$\lim_{\varepsilon \downarrow 0} \operatorname{dist} \{P(t, \cdot; \varepsilon), \pi_p \cdot \varphi(\cdot)\} = 0 \quad \text{for } 0 \leqq t \leqq t_\varepsilon,$$

$$\lim_{\varepsilon \downarrow 0} \varepsilon t_\varepsilon = \tau_0.$$

THEOREM 3.5. *Suppose that π_p is a nondegenerate periodic orbit and that $P(t, x; \varepsilon)$ is a periodic solution of $(2.1)_\varepsilon$, (2.2) associated with π_p as shown in Theorem 3.4. Let $U_\varepsilon(t)$ be a period map of $P(t, x; \varepsilon)$ in $(2.1)_\varepsilon$, (2.2). Then, there exist $\varepsilon_1 > 0$ and $C_i > 0$ $(i = 1, 2, 3)$ such that the spectrum of $U_\varepsilon(t)$ consists of two spectral sets $\sigma_1(\varepsilon)$, $\sigma_2(\varepsilon)$ satisfying $\sigma_1(\varepsilon) \subset \{\lambda \in \mathbb{C} \,|\, C_1 \leqq |\lambda| \leqq C_2\}$ and $\sigma_2(\varepsilon) \subset \{\lambda \in \mathbb{C} \,|\, |\lambda| \leqq C_3 \varepsilon\}$ for $0 < \varepsilon \leqq \varepsilon_1$. For any $\lambda(\varepsilon) \in \sigma_1(\varepsilon)$, there is some characteristic multiplier λ_0 of (3.1) such that $\lim_{\varepsilon \downarrow 0} \lambda(\varepsilon) = \lambda_0$ and its converse also holds.*

Remark 3.1. The spectrum of $U_\varepsilon(t)$ is independent of t because of the compactness of $U_\varepsilon(t)$.

THEOREM 3.6. *Assume π_p is nondegenerate and the spectrum of the period map of (3.1) (except 1) lies in $\{|\lambda| < \delta\}$ for some $\delta < 1$. Then $P(t, x; \varepsilon)$ associated with π_p is*

orbitally asymptotically stable for sufficiently small $\varepsilon > 0$. Moreover, it has an attractive region V that is independent of ε, in $C^0(\Omega; \mathbb{R}^n)$.

These results imply that asymptotic states of solutions of PDE problems derive from those of the approximating ODE problems except for degenerate cases. Consequently, the asymptotic behavior of solutions of the PDE problem can be generically investigated by studying the ODE problems.

Let θ be a parameter indicating the heterogeneities of circumstances for individuals in ecology. Then problem $(2.1)_\varepsilon$, (2.2) is written as

$$\frac{\partial u}{\partial t} + \operatorname{div} J(u; \theta) = \varepsilon f(x, u; \theta), \quad t > 0, \quad x \in \Omega, \tag{3.2}_\varepsilon$$

$$\langle \nu, J(u; \theta) \rangle = 0_n, \quad t > 0, \quad x \in \partial\Omega, \tag{3.3}$$

and the stationary solution $\varphi(x)$ of (2.5) as $\varphi(x; \theta)$.

The approximating ODE becomes

$$\frac{dU}{d\tau} = F(U; \theta), \qquad \tau > 0 \tag{3.4}$$

with a parameter θ. For a study of pattern formation due to spatial inhomogeneity, we are interested in bifurcation problems of $(3.2)_\varepsilon$, (3.3) with respect to θ. Suppose that solution structures of (3.4) can be globally analyzed with respect to θ. Then Theorems 3.1–3.6 imply that the regular (i.e., nondegenerate) branch for θ of (3.4) extends to that of the PDE problem $(3.2)_\varepsilon$, (3.3). However, we should note that these theorems do not give any information to degenerate cases such as bifurcation points or limiting points. It is expected that the global picture of asymptotic states for PDE problems will be generically deformed from that of associated ODE problems. These imperfection problems are now in progress.

4. Proofs. Let B be the Banach space $C^0(\Omega; \mathbb{R}^n)$ with sup-norm $\|\cdot\|_0$ and let A be the operator defined by $Au = -\operatorname{div} J(u)$ with domain $D(A) = \{u \in W^{2,N+1}(\Omega; \mathbb{R}^n) \mid Au \in B \text{ and } \langle \nu, J(u) \rangle = 0_n \text{ on } x \in \partial\Omega\}$. $\mathscr{L}(B)$ denotes the Banach space of bounded linear operators from B into itself with operator norm $\|\cdot\|$. Then A is a generator of an analytic semigroup in B with the spectrum $\sigma(A) = \{0 = \lambda_0 > \lambda_1 > \lambda_2 > \cdots\}$ (Stewart [36]), $\operatorname{Ker} A = \mathbb{R}^n \cdot \varphi$, and the projection $Q: B \to \operatorname{Ker} A$ is given by $Qu = \int_\Omega u(x)\, dx \cdot \varphi$ for $u \in B$. Under this notation, $F(U)$ in $(2.9)_1$ is represented by

$$F(U) \cdot \varphi(\cdot) = Qf(\cdot, U \cdot \varphi(\cdot)) \quad \text{for } U \in \mathbb{R}^n \tag{4.1}$$

and problem $(2.1)_\varepsilon$, (2.2), (2.3) is written as the initial value problem in B:

$$\begin{aligned} &\frac{du}{dt} = Au + \varepsilon f(u), \qquad t > 0, \\ &u(0) = \xi, \end{aligned} \tag{4.2}$$

where $f(u)(\cdot)$ denotes $f(\cdot, u)$, so that $f(u)$ is a smooth function from B into itself. Here we give other notation used in this section: $\sigma(\mathscr{L})$ denotes the spectrum of $\mathscr{L}$ if $\mathscr{L}$ is a closed operator in B; Id is the identity in B; $P = \mathrm{Id} - B$; $B_1 = QB$; $B_2 = PB$; I_i is the identity in B_i $(i = 1, 2)$, respectively; E is a unit matrix in $\mathbb{R}^n$; M and a are constants such that $\|e^{tA}P\| \leqq M e^{-at}$ for $t \geqq 0$; C and C_i $(i = 1, 2, \cdots)$ are constants independent of sufficiently small $\varepsilon > 0$.

Proof of Theorem 3.1. The stationary problem (2.1), (2.2) is written as

$$0 = Au + \varepsilon f(u). \tag{4.3}$$

We define u_1 and u_2 by $u_1 = Qu$ and $u_2 = Pu$, and we decompose (4.3) into

$$0 = Qf(u_1+u_2), \qquad 0 = Au_2 + \varepsilon Pf(u_1+u_2). \tag{4.4}$$

By (4.1), the nondegeneracy of π_e is equivalent to $\det (Qf_u(\pi_e \cdot \varphi)|_{B_1}) \neq 0$, so that by the standard Lyapunov-Schmidt method, we see that (4.4) has a solution $(\bar{u}_1(\varepsilon), \bar{u}_2(\varepsilon))$ with $(\bar{u}_1(0), \bar{u}_2(0)) = (\pi_e \cdot \varphi, 0)$ (see Theorem 3 of Ei and Mimura [5], for instance). □

Proof of Theorem 3.2. Let $\bar{u}(\varepsilon)(\cdot) = \bar{u}(\cdot\,; \varepsilon)$ $(|\varepsilon| \leqq \varepsilon_0)$ be the stationary solution of (4.3) satisfying $\bar{u}(0)(\cdot) = \pi_e \cdot \varphi(\cdot)$. Then the linearized operator L_ε of (4.2) about $\bar{u}(\varepsilon)$ is described by $L_\varepsilon = A + \varepsilon f_u(\bar{u}(\varepsilon))$. Consider the equation

$$(\lambda - L_\varepsilon)u = v \tag{4.5}$$

for $\lambda \in \mathbb{C}$ and $v \in B$. Putting $u_1 = Qu$, $u_2 = Pu$ and $v_1 = Qv$, $v_2 = Pv$, we have a pair of equations equivalent to (4.5):

$$\lambda u_1 - \varepsilon Qf_u(\bar{u}(\varepsilon))(u_1+u_2) = v_1, \tag{4.6}_1$$

$$\lambda u_2 - A_2 u_2 - \varepsilon Pf_u(\bar{u}(\varepsilon))(u_1+u_2) = v_2, \tag{4.6}_2$$

where A_2 is the operator that is restricted to B_2, which is invertible in B_2. Let α be a constant satisfying $0 > -\alpha > \lambda_1$. Then, if λ satisfies $\operatorname{Re}\lambda > -\alpha$ and $\varepsilon > 0$ is sufficiently small, $(4.6)_2$ is solvable on u_2 and it is written as

$$u_2 = G(\lambda, \varepsilon)(v_2 + \varepsilon Pf_u(\bar{u}(\varepsilon))), \tag{4.7}$$

where $G_1(\lambda, \varepsilon) = (\lambda I_2 - A_2 - \varepsilon Pf_u(\bar{u}(\varepsilon)))^{-1}$. Substituting (4.7) into $(4.6)_1$, we have

$$\{\lambda I_1 - \varepsilon Qf_u(\bar{u}(\varepsilon))(I_1 + \varepsilon G(\lambda, \varepsilon)Pf_u(\bar{u}(\varepsilon)))\}u_1 = v_1 + \varepsilon Qf_u(\bar{u}(\varepsilon))G(\lambda, \varepsilon)v_2. \tag{4.8}$$

Since $\bar{u}(\varepsilon)$ and $G_1(\lambda, \varepsilon)$ are uniformly bounded for sufficiently small $\varepsilon > 0$, there is a constant $C_0 > 0$ such that $\|Qf_u(\bar{u}(\varepsilon))(I_1 + \varepsilon G_1(\lambda, \varepsilon)Pf_u(\bar{u}(\varepsilon)))\| \leqq C_0$. Hence, if λ satisfies $|\lambda| > C_1\varepsilon$ for some $C_1 > C_0$ and sufficiently small $\varepsilon > 0$, (4.8) is solvable on u_1. Consequently, there are positive constants C_1 and ε_1 such that if λ satisfies $|\lambda| \geqq C_1\varepsilon$ and $\operatorname{Re}\lambda > -\alpha$ for $0 < \varepsilon \leqq \varepsilon_1$, λ is an element of $\rho(L_\varepsilon)$. This implies that the spectrum $\sigma(L_\varepsilon)$ consists of two spectral sets $\sigma_1(\varepsilon)$, $\sigma_2(\varepsilon)$ such that $\sigma_1(\varepsilon) \subset \{\lambda \in \mathbb{C} \,|\, |\lambda| \leq C_1\varepsilon\}$ and $\sigma_2(\varepsilon) \subset \{\lambda \in \mathbb{C} \,|\, \operatorname{Re}\lambda < -\alpha\}$.

We consider the spectrum contained in $\sigma_1(\varepsilon)$ for $0 < \varepsilon \leqq \varepsilon_1$. Since L_ε has the compact resolvent, $\sigma_1(\varepsilon)$ consists only of isolated eigenvalues with finite multiplicity. Putting $\lambda' = \lambda/\varepsilon$ for $\lambda \in \sigma_1(\varepsilon)$, we consider the following eigenvalue problem:

$$(\varepsilon\lambda' - L_\varepsilon)u = 0,$$

which is equivalent to

$$\lambda' u_1 - Qf_u(\bar{u}(\varepsilon))(u_1+u_2) = 0, \tag{4.9}_1$$

$$\varepsilon\lambda' u_2 - A_2 u_2 - \varepsilon Pf_u(\bar{u}(\varepsilon))(u_1+u_2) = 0, \tag{4.9}_2$$

where $u_1 = Qu$ and $u_2 = Pu$. Since $(4.9)_2$ is solvable on u_2, we obtain

$$u_2 = \varepsilon G_2(\lambda', \varepsilon)Pf_u(\bar{u}(\varepsilon))u_1, \tag{4.10}$$

where $G_2(\lambda', \varepsilon) = (\varepsilon\lambda' I_2 - A_2 - \varepsilon Pf_u(\bar{u}(\varepsilon)))^{-1}$. After substituting (4.10) into $(4.9)_1$, we see $G_3(\lambda', \varepsilon)u_1 = 0$, where $G_3(\lambda', \varepsilon) = \{\lambda' I_1 - Qf_u(\bar{u}(\varepsilon))(I_1 + \varepsilon G_2(\lambda', \varepsilon)Pf_u(\bar{u}(\varepsilon)))\}$. Here $G_3(\lambda', \varepsilon)$ is a matrix in the finite-dimensional space B_1, so that it suffices to investigate zeros of $g_\varepsilon(\lambda') = \det G_3(\lambda', \varepsilon)$. Let $D_1 = \{\lambda \in \mathbb{C} \,|\, |\lambda| < C_1\}$. $g_\varepsilon(\lambda')$ is an analytic function of λ', and as $\varepsilon \downarrow 0$, it converges to $g_0(\lambda') = \det(\lambda' I_1 - Qf_u(\pi_e \cdot \varphi)|_{B_1}) = \det(\lambda' E - F_U(\pi_e))$ uniformly on $\bar{D}_1$. We may assume that $g_\varepsilon(\lambda')$ and $g_0(\lambda')$ are not zero on ∂D_1, so that the proof is complete by use of the theorem of Hurwitz. □

Proof of Theorem 3.3. From Theorem 3.2, it is obvious that $\bar{u}(\varepsilon)$ is asymptotically stable for sufficiently small $\varepsilon > 0$. We show only that $\bar{u}(\varepsilon)$ has an attractive region independent of small $\varepsilon > 0$.

Under the assumption of Theorem 3.3, it is proved by Lemma 5.1 of Ei and Mimura [5] that there exist positive constants β, θ $(0 < \theta < \pi/2)$, ε_2, and C such that for $0 < \varepsilon < \varepsilon_2$, the sector $S_\varepsilon = \{\lambda \in \mathbb{C} \,|\, |\arg(\lambda - \beta\varepsilon)| < \pi/2 + \theta,\ \lambda \neq \beta\varepsilon\}$ is contained in $\rho(L_\varepsilon)$ and

$$\|(\lambda - L_\varepsilon)^{-1}\| \leqq \frac{C}{|\lambda - \beta\varepsilon|} \quad \text{for all } \lambda \in S_\varepsilon \tag{4.11}$$

holds. By transforming t into τ $(=\varepsilon t)$, (4.2) becomes

$$\frac{du}{d\tau} = \frac{1}{\varepsilon} Au + f(u) \tag{4.12}$$

and the linearized operator about $\bar{u}(\varepsilon)$ becomes $1/\varepsilon L_\varepsilon$ $(=1/\varepsilon A + f_u(\bar{u}(\varepsilon)))$. Let $S = \{\lambda \in \mathbb{C} \,|\, |\arg(\lambda - \beta)| < \pi/2 + \theta,\ \lambda \neq \beta\}$. Then by (4.11), we have $\|(\lambda - 1/\varepsilon L_\varepsilon)^{-1}\| \leqq C/|\lambda - \beta|$ for all $\lambda \in S$, implying that $\|e^{\tau L_\varepsilon/\varepsilon}\| \leqq C_2 e^{-\beta\tau}$ for some $C_2 > 0$ independent of ε satisfying $0 < \varepsilon \leqq \varepsilon_2$. Thus it turns out that $\bar{u}(\varepsilon)$ of (4.12) has an attractive region V independent of sufficiently small $\varepsilon > 0$. □

Proof of Theorem 3.4.

LEMMA 4.1. *For a given constant $C_0 > 0$, there is $\varepsilon_0 > 0$ such that* (4.2) *has an invariant manifold $\mathcal{M}_\varepsilon$ for $|\varepsilon| < \varepsilon_0$ which is represented as $\mathcal{M}_\varepsilon = \{u_1 + h(u_1, \varepsilon) \,|\, u_1 \in D_0\}$, with $D_0 = \{u_1 \in B_1 \,|\, |u_1| < C_0\}$. h satisfies $h(\cdot, \cdot) \in C^1(D_0 \times (-\varepsilon_0, \varepsilon_0); B_2)$, $\|h(u_1, \varepsilon)\| \leqq C_1\varepsilon$, and $\|h_u(u_1, \varepsilon)\| \leqq C_1\varepsilon$ for some $C_1 > 0$.*

Proof. Problem (4.2) is rewritten as follows:

$$\frac{du_1}{dt} = \varepsilon Q f(u_1 + u_2), \qquad \frac{du_2}{dt} = A_2 u_2 + \varepsilon P f(u_1 + u_2), \tag{4.13}$$

where $u_1 = Qu$ and $u_2 = Pu$. Noting $\|e^{tA_2}\| = \|e^{tA} P\| \leqq M e^{-at}$, we can prove this lemma in a standard manner (cf. Carr [2]), so we omit the proof. □

By Lemma 4.1, we find that the dynamics of (4.13) on the invariant manifold $\mathcal{M}_\varepsilon$ is reduced to that of $du_1/dt = \varepsilon Q f(u_1 + h(u_1, \varepsilon))$, or equivalently,

$$\frac{du_1}{d\tau} = Q f(u_1 + h(u_1, \varepsilon)), \tag{4.14}_\varepsilon$$

where $\tau = \varepsilon t$. Since $\pi_e(\tau) \cdot \varphi(\cdot)$ is a nondegenerate periodic solution of $(4.14)_0$ ($\varepsilon = 0$ in $(4.14)_\varepsilon$) with period τ_0, it is shown by Theorem 8.3.2 of Henry [18] that there exists a periodic solution $\tilde{P}(\tau; \varepsilon)$ with period τ_ε for sufficiently small ε such that $\tau_\varepsilon \to \tau_0$ and $\operatorname{dist}\{\tilde{P}(\tau; \varepsilon)(\cdot), \pi_p \cdot \varphi(\cdot)\} \to 0$ as $\varepsilon \downarrow 0$. Therefore the definition of $P(t, \cdot; \varepsilon) = \tilde{P}(\varepsilon t; \varepsilon)(\cdot)$ completes the proof. □

Remark 4.1. We note that $(4.14)_\varepsilon$ is reduced to $(2.9)_1$ as $\varepsilon \downarrow 0$. This implies that $(2.9)_1$ constructed by the two-timing method is the lowest approximation to $(4.14)_\varepsilon$ on the invariant manifold $\mathcal{M}_\varepsilon$.

Proof of Theorem 3.5. We define $P(t, \cdot; \varepsilon)$ by $P(t; \varepsilon)(\cdot)$ and consider the linearized equation of (4.2) about $P(t; \varepsilon)$:

$$\frac{du}{dt} = (A + \varepsilon f_u(P(t; \varepsilon)))u. \tag{4.15}$$

Let $T_\varepsilon(t, s)$ be the evolutional operator of (4.15), that is, $T_\varepsilon(t, s)u_0$ gives a solution of (4.15) with $u(s) = u_0$. Then the period map $U_\varepsilon(s)$ of $P(t; \varepsilon)$ is given by $U_\varepsilon(s) = T_\varepsilon(s + t_\varepsilon, t_\varepsilon)$.

LEMMA 4.2. *There exist* $\varepsilon_1>0$ *and* $C_1>0$ *such that* $\|PU_\varepsilon(s)\|\leqq C_1\varepsilon$ *for* $0<\varepsilon<\varepsilon_1$ *and* $s\geqq 0$.

Proof. Equation (4.15) is equivalent to

$$\frac{du_1}{dt}=\varepsilon Qf_u(P(t;\varepsilon))(u_1+u_2), \tag{4.16}_1$$

$$\frac{du_2}{dt}=A_2u_2+\varepsilon Pf_u(P(t;\varepsilon))(u_1+u_2), \tag{4.16}_2$$

where $u_1=Qu$ and $u_2=Pu$. Let $u(s)=u_0$ with $\|u_0\|_0\leqq 1$. Then, as is done in Lemma 6.6 of Ei [4], we can show that there exist $\varepsilon_2>0$ and $C_2>0$ such that $\|u_2(s+t)-e^{tA_2}Pu_0\|_0\leqq C_2\varepsilon$ for $0<\varepsilon<\varepsilon_2$, so that we have

$$\|u_2(s+t)\|_0\leqq C_2\varepsilon+Me^{-at}\|u_0\|_0. \tag{4.17}$$

Since the inequality $t_\varepsilon\geqq C_3/\varepsilon$ holds for some $C_3>0$, we see that

$$\|u_2(s+t_\varepsilon)\|_0=\|PU_\varepsilon u_0\|_0\leqq C_2\varepsilon+Me^{-aC_3/\varepsilon}\leqq C_4\varepsilon$$

for some $C_4>0$. The estimate above holds uniformly for u_0 ($\|u_0\|_0\leqq 1$), which gives the proof. □

Suppose $\|PU_\varepsilon(s)\|\leqq C_1\varepsilon$ for $0<\varepsilon<\varepsilon_1$ and consider the equation $(\lambda-U_\varepsilon(s))u=v$, or equivalently

$$\lambda u_1-QU_\varepsilon(s)(u_1+u_2)=v_1, \tag{4.18}_1$$

$$\lambda u_2-PU_\varepsilon(s)(u_1+u_2)=v_2, \tag{4.18}_2$$

where $u_1=Qu$, $u_2=Pu$ and $v_1=Qv$, $v_2=Pv$. If λ satisfies $|\lambda|>C_2\varepsilon$ for some C_2 with $C_2>C_1$, then $(4.18)_2$ is solvable on u_2 and we have

$$u_2=(\lambda-PU_\varepsilon(s))^{-1}(v_2-PU_\varepsilon(s)u_1), \tag{4.19}$$

$$\|(\lambda-PU_\varepsilon(s))^{-1}\|\leqq\frac{C_3}{|\lambda|} \tag{4.20}$$

for some $C_3>0$. Substituting (4.19) into $(4.18)_1$, we see

$$(\lambda-QU_\varepsilon(s)Q+QU_\varepsilon(s)(\lambda-PU_\varepsilon(s))^{-1}PU_\varepsilon(s))u_1=k_1, \tag{4.21}$$

where $k_1=v_1-QU_\varepsilon(s)(\lambda-PU_\varepsilon(s))^{-1}v_2$. Let $\Pi(\tau)$ be the periodic map of π_p in

$$\frac{du_1}{d\tau}=Qf_u(\pi_p(\tau)\cdot\varphi)u_1 \quad \text{for } u_1\in B_1,$$

which is the same equation as (3.1). Since $\|QU_\varepsilon(s)Q-\pi(\varepsilon s)\|\to 0$ as $\varepsilon\downarrow 0$, there exist C_i $(i=4,5,6)$ such that $\sigma(QU_\varepsilon(s)Q)\subset\{\lambda\in\mathbb{C}\,|\,C_4<|\lambda|<C_5\}$ and such that, for any λ satisfying $|\lambda|\leqq C_4$ or $|\lambda|\geqq C_5$, $(\lambda-QU_\varepsilon(s)Q)^{-1}$ exists with $\|(\lambda-QU_\varepsilon(s)Q)^{-1}\|\leqq C_6$. Hence there exists $C_7>0$ such that

$$\|(\lambda-QU_\varepsilon(s)Q)^{-1}QU_\varepsilon(s)(\lambda-PU_\varepsilon(s))^{-1}PU_\varepsilon(s)\|\leqq\frac{C_7\varepsilon}{|\lambda|}\leqq\frac{C_7}{C_2}$$

for λ satisfying $C_2\varepsilon<|\lambda|\leqq C_4$ or $|\lambda|\geqq C_5$. Taking C_2 such that $C_2>C_7$, we see that (4.21) is solvable on u_1 when $\varepsilon>0$ is sufficiently small, that is, $\{\lambda\in\mathbb{C}\,|\,C_2\varepsilon<|\lambda|\leqq C_4$ or $|\lambda|\geqq C_5\}$ is in $\rho(U_\varepsilon(s))$, which gives the proof of the first half of this theorem.

LEMMA 4.3. $\|U_\varepsilon(s)-\pi(\varepsilon s)Q\|\to 0$ *as* $\varepsilon\downarrow 0$.

Proof. From (4.17), we suppose that there exists a positive constant C_8 such that

$$\|u_2(s+t)\|_0\leqq C_8(\varepsilon+e^{-at}) \tag{4.22}$$

uniformly for u_0 with $\|u_0\|_0 \leqq 1$. Let $\bar{T}_\varepsilon(t, s)$ be the evolutional operator of

$$\frac{du}{dt} = \varepsilon Q f_u(P(t; \varepsilon))u \tag{4.23}$$

and let $\bar{U}_\varepsilon(s)$ be the period map, that is, $\bar{U}_\varepsilon(s) = \bar{T}_\varepsilon(s + t_\varepsilon, s)$. Then $\|\bar{T}_\varepsilon(t, s)\|$ is bounded for $t_\varepsilon \geqq t \geqq s \geqq 0$ and $\bar{U}_\varepsilon(s)Q \to \pi(\varepsilon s)Q$ as $\varepsilon \downarrow 0$. Equation $(4.16)_1$ implies that

$$u_1(s + t_\varepsilon) = \bar{U}_\varepsilon(s)Qu_0 + \varepsilon \int_s^{s+t_\varepsilon} \bar{T}_\varepsilon(s + t_\varepsilon, \sigma)Qf_u(P(\sigma; \varepsilon))u_2(\sigma)\, d\sigma. \tag{4.24}$$

From (4.22) we have

$$\begin{aligned} \|u_1(s + t_\varepsilon) - \bar{U}_\varepsilon(s)Qu_0\|_0 &= \|QU_\varepsilon(s)u_0 - \bar{U}_\varepsilon(s)Qu_0\|_0 \\ &\leqq \varepsilon \int_s^{s+t_\varepsilon} C_9(\varepsilon + e^{-a\sigma})\, d\sigma \\ &\leqq \varepsilon C_9\left(\varepsilon t_\varepsilon + \frac{1}{a}\right) \\ &\leqq \varepsilon C_{10} \end{aligned} \tag{4.25}$$

for some C_9 and C_{10}. Hence

$$\begin{aligned} \|U_\varepsilon(s) - \Pi(\varepsilon s)Q\| &\leqq \|QU_\varepsilon(s) - \bar{U}_\varepsilon(s)Q\| \\ &\quad + \|\bar{U}_\varepsilon(s)Q - \Pi(\varepsilon s)Q\| + \|PU_\varepsilon(s)\| \to 0 \end{aligned} \tag{4.26}$$

as $\varepsilon \downarrow 0$. □

We consider the spectrum contained in $\sigma_1(\varepsilon)$. Noting that $\sigma(U_\varepsilon(s)) = \sigma_1(\varepsilon) \cup \sigma_2(\varepsilon)$ such that $\sigma_1(\varepsilon) \subset \{\lambda \in \mathbb{C} \mid C_4 < |\lambda| < C_5\}$ and $\sigma_2(\varepsilon) \subset \{\lambda \in \mathbb{C} \mid |\lambda| \leqq C_2\varepsilon\}$, we define a projection corresponding to $\sigma_1(\varepsilon)$ by $Q_\varepsilon(s)$, that is, $Q_\varepsilon(s) = 1/2\pi i \int_\Gamma (\lambda - U_\varepsilon(s))^{-1}\, d\lambda$, where Γ is a closed curve enclosing $\sigma_1(\varepsilon)$ with $C_2\varepsilon < |\gamma| < C_4$ or $|\gamma| > C_5$ for any $\gamma \in \Gamma$. Let $B_\varepsilon(s) = Q_\varepsilon(s)B$ and $I_\varepsilon(s)$ be the identity in $B_\varepsilon(s)$. Since $U_\varepsilon(s)$ has a compact resolvent, $\sigma_1(\varepsilon)$ consists only of isolated eigenvalues with finite multiplicity. Hence $B_\varepsilon(s)$ is a finite-dimensional space, and it suffices to investigate the zeros of $g_{\varepsilon,s}(\lambda) = \det(\lambda I_\varepsilon(s) - Q_\varepsilon(s)U_\varepsilon(s))$. From Lemma 4.3 and $Q = 1/2\pi i \int_\Gamma (\lambda - \pi(\varepsilon s)Q)^{-1}\, d\lambda$, we have $\|Q_\varepsilon(s) - Q\| \to 0$ as $\varepsilon \downarrow 0$, so that $g_{\varepsilon,s}(\lambda)$ converges to $g_0(\lambda) = \det(\lambda I_1 - Q\Pi(0)Q) = \det(\lambda I_1 - \Pi(0))$ uniformly for λ satisfying $C_4 \leqq |\lambda| \leqq C_5$. Since $g_{\varepsilon,s}(\lambda)$ and $g_0(\lambda)$ are analytic functions of λ and $g_0(\lambda) \neq 0$ on Γ, the theorem of Hurwitz completes the proof. □

Proof of Theorem 3.6. If we transform t into $\tau\ (= \varepsilon t)$, (4.15) becomes

$$\frac{du}{d\tau} = \left(\frac{1}{\varepsilon}A + \varepsilon f_u(\tilde{P}(\tau; \varepsilon))\right)u, \tag{4.27}$$

where $\tilde{P}(\tau; \varepsilon)$ is a periodic solution with period τ_ε associated with $\pi_p \cdot \varphi$ as shown in the proof of Theorem 3.4. Let $\tilde{T}_\varepsilon(\tau, s)$ and $\tilde{U}_\varepsilon(s)$ be the evolutional operator and the period map of (4.27), respectively. Then we remark that $\tilde{T}_\varepsilon(\varepsilon t, \varepsilon s) = T_\varepsilon(t, s)$ and $\tilde{U}_\varepsilon(\varepsilon s) = U_\varepsilon(s)$ hold, where $T_\varepsilon(t, s)$ and $U_\varepsilon(s)$ are the evolutional operator and the period map of (4.15), respectively. From (4.22) and (4.25), we see that there exists $C_{11} > 0$ such that

$$\|\tilde{T}(\tau, s)\| \leqq C_{11} \tag{4.28}$$

for all $\tau \geqq s \geqq 0$. Then, in a way similar to the proof of Theorem 7.2.3 of Henry [18], we can prove that there exist $C_{12} > 0$ and $\delta > 0$ independent of sufficiently small $\varepsilon > 0$ such that

$$\|\tilde{T}(\tau, s)u\|_0 \leqq C_{12}\, e^{-\delta(\tau - s)}\|u\|_0 \tag{4.29}$$

for $\tau \geqq s$ and $u \in B_\varepsilon^1(s)$, where $B_\varepsilon^1(s)$ is a subspace invariant under $\tilde{U}(s)$, $\sigma(\tilde{U}(s)|_{B_\varepsilon^1}) = \sigma(\tilde{U}(s))\backslash\{1\}$. The estimates (4.28) and (4.29) imply this proof. □

5. Application to two-competing species models. In the previous section, we have found that when ε is sufficiently small, the qualitative property of solutions of PDE problem $(2.1)_\varepsilon$, (2.2) is similar to those of ODE problem $(2.9)_1$, if they are nondegenerate.

In this section, as an application of our procedure, we consider the two-competing-species model already proposed as $(1.5)_\varepsilon$, (1.6) in § 1. For this model, we introduced a parameter θ in some interval Θ indicating the distance between the maximum points of $K_1(x)$ and $K_2(x)$. Here, to make clear the dependence on θ, we write the carrying capacity $K(x)$ as $K(x; \theta)$ and the stationary solution $\varphi(x)$ of (2.5) as $\varphi(x; \theta)$. Then, applying our procedure to $(1.5)_\varepsilon$, we obtain the following approximating ODE with the parameter θ:

$$\begin{aligned} \frac{dU_1}{d\tau} &= (r_1 - a_1(\theta)U_1 - b_1(\theta)U_2)U_1, \\ \frac{dU_2}{d\tau} &= (r_2 - b_2(\theta)U_1 - a_2(\theta)U_2)U_2, \end{aligned} \tag{5.1}$$

where

$$a_i(\theta) = \alpha_i \int_\Omega \frac{\varphi_i^2(x; \theta)}{K_i(x; \theta)}\,dx, \qquad b_i(\theta) = \beta_i \int_\Omega \frac{\varphi_1(x; \theta)\varphi_2(x; \theta)}{K_i(x; \theta)}\,dx$$

for $i = 1, 2$. The asymptotic states of solutions to (5.1) can be studied by phase-plane analysis, namely, the global structure of equilibria of (5.1) is completely known with respect to θ.

Let us show one example. We choose

$$\begin{aligned} K_1(x; \theta) &= \begin{cases} \dfrac{3}{5} \cdot \cos 4\pi\left(x + \dfrac{\theta}{2}\right) + 1 & \left(-\dfrac{\theta}{2} + \dfrac{1}{4} < x < -\dfrac{\theta}{2} + \dfrac{3}{4}\right), \\ \dfrac{2}{5} & \text{otherwise}, \end{cases} \\ K_2(x; \theta) &= \begin{cases} \dfrac{3}{5} \cdot \cos 4\pi\left(x - \dfrac{\theta}{2}\right) + 1 & \left(\dfrac{\theta}{2} + \dfrac{1}{4} < x < \dfrac{\theta}{2} + \dfrac{3}{4}\right), \\ \dfrac{2}{5} & \text{otherwise}. \end{cases} \end{aligned} \tag{5.2}$$

By noting that $x_1 = -(\theta/2) + \frac{1}{2}$ and $x_2 = \theta/2 + \frac{1}{2}$, we find that θ varies in the interval $\Theta = [0, \frac{1}{2}]$. We specify d, r, α, β, and γ as

$$\begin{aligned} &d = (d_1, d_2) = (1, 1), \quad r = (1, 1), \quad \alpha = (1, 1), \\ &\beta = (1.5, 1.2), \qquad \gamma = 1.6, \end{aligned} \tag{5.3}$$

respectively.

First we consider the case when the dispersal is ignored, so that $(1.5)_\varepsilon$ is formally reduced to the ODE system with parameters x and θ:

$$\begin{aligned} \frac{du_1}{dt} &= \varepsilon\left(r_1 - \frac{\alpha_1 u_1 + \beta_1 u_2}{K_1(x; \theta)}\right)u_1, \\ \frac{du_2}{dt} &= \varepsilon\left(r_2 - \frac{\beta_1 u_1 + \alpha_2 u_2}{K_2(x; \theta)}\right)u_2. \end{aligned} \tag{5.4}$$

A simple calculation shows that the parameters in (5.4) are chosen so as to have *no* stable coexistence equilibrium for any $x \in \bar{\Omega}$ and $\theta \in \Theta$. In ecological terms, this situation indicates the competitive exclusion between the two species as long as they do not migrate. Under this situation, we have the following problem. If they migrate in a heterogeneous habitat, is it possible for them to coexist? For this problem, the approximating ODE system (5.1) works well.

First, consider the case when the advection is ignored, that is, $\gamma = 0$. We easily find that the stationary solutions of $(1.5)_\varepsilon$ with $\varepsilon = 0$ and of (1.6) are $\varphi(x; \theta) \equiv (1, 1)$, which is independent of θ, so that the approximating ODE system is also independent of θ. Phase-plane analysis shows the global picture of equilibria of (5.1) in Fig. 3. The ecological interpretation is that if the two competing species move with diffusion only, they never coexist even if the heterogeneous habitat encourages a favorably segregated pattern for them.

Next, consider the case when the advection term is present, say $\gamma = 1.6$. Figure 4 shows the global bifurcation pictures for θ, where there are two critical values θ_* ($\simeq 0.15$) and θ^* ($\simeq 0.27$) such that *three* qualitatively different asymptotic states appear for $0 < \theta < \theta_*$, $\theta_* < \theta < \theta^*$, and $\theta_* < \theta < \frac{1}{2}$. The stable coexistence equilibrium exists for $\theta_* < \theta < \frac{1}{2}$. Theorems 3.1-3.3 say that there is a stable stationary coexistence solution of $(1.5)_\varepsilon$, (1.6). That is, coexistence of the competing species occurs due to incorporation of spatial heterogeneity and tactical migration (see [34]). Thus, from Fig. 4, the reader can now completely understand why three different asymptotic states of solutions appear for suitable θ as in Fig. 2.

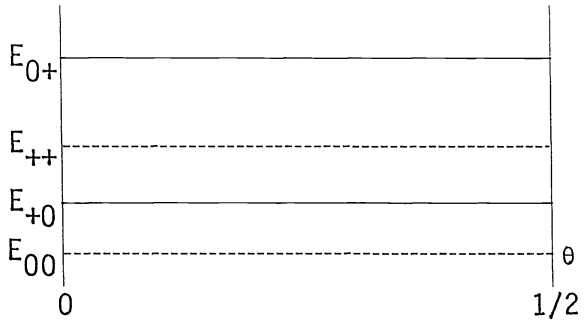


FIG. 3. *Global structure of equilibria of* (4.4) *when* $\gamma = 0$. $E_{00} = (0, 0)$, $E_{+0} = (r_1/a_1, 0)$, $E_{0+} = (0, r_2/a_2)$, *and* E_{++} *is the coexistence equilibrium.* —— *is the stable branch.* - - - *is the unstable branch.*

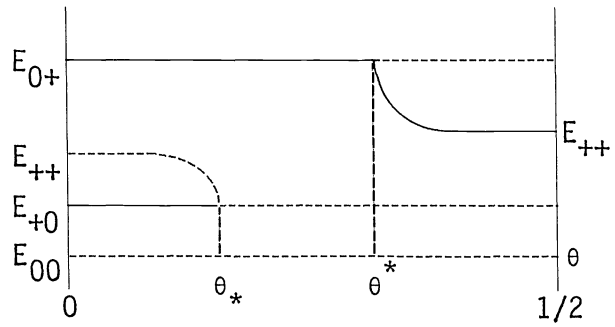


FIG. 4. *Global structure of equilibria of* (4.4) *when* $\gamma = 1.6$.

There remains one interesting problem. Our theorems do not discuss bifurcation points $\theta = \theta_*$ and $\theta = \theta^*$. This analysis, with further applications to the multispecies model, will be reported in [6].

Acknowledgment. We are grateful to Miss T. Yamanoue for numerically computing the problems $(1.5)_\varepsilon$, (1.6) with (5.2) and (5.3).

REFERENCES

[1] S. B. ANGENENT, J. MALLET-PARET, AND L. A. PELETIER, *Stable transition layers in a semilinear boundary value problem*, J. Differential Equations, 67 (1987), pp. 212-242.

[2] J. CARR, *Application of Centre Manifold Theory*, Appl. Math. Sci. 35, Springer-Verlag, Berlin, New York, 1981.

[3] E. CONWAY, D. HOFF, AND J. SMOLLER, *Large time behavior of solutions of systems of nonlinear reaction-diffusion equations*, SIAM J. Appl. Math., 35 (1978), pp. 1-16.

[4] S.-I. EI, *Two-timing methods with applications to heterogeneous reaction-diffusion system*, Hiroshima Math. J., 18 (1988), pp. 127-160.

[5] S.-I. EI AND M. MIMURA, *Transient and large time behaviors of solutions to heterogeneous reaction-diffusions*, Hiroshima Math. J., 14 (1985), pp. 649-678.

[6] S.-I. EI, M. MIMURA, AND T. YAMANOUE, *Pattern formation due to spatial heterogeneities in population models*, manuscript.

[7] P. C. FIFE AND L. A. PELETIER, *Clines Induced by Variable Migration*, Lecture Notes in Biomath. 38, Springer-Verlag, Berlin, New York, 1980, pp. 276-278.

[8] W. H. FLEMING, *A selection-migration model in population genetics*, J. Math. Biol., 2 (1975), pp. 219-233.

[9] C. FOIAS, G. R. SELL, AND R. TEMAM, *Inertial manifolds for nonlinear evolutionary equations*, J. Differential Equations, 73 (1988), pp. 309-353.

[10] G. FUSCO, *On the explicit construction of an* ODE *which has the same dynamics as a scalar parabolic* PDE, J. Differential Equations, 69 (1987), pp. 85-110.

[11] G. FUSCO AND J. K. HALE, *Stable equilibria in a scalar parabolic equation with variable diffusion*, SIAM J. Math. Anal., 16 (1985), pp. 152-164.

[12] W. S. C. GURNEY AND R. M. NISBET, *The regulation of inhomogeneous populations*, J. Math. Biol., 52 (1975), pp. 441-457.

[13] J. K. HALE, *Large diffusivity and asymptotic behavior in parabolic systems*, J. Math. Anal. Appl., 118 (1986), pp. 455-466.

[14] J. K. HALE AND M. CHIPOT, *Stable equilibria with variable diffusion*, Contemp. Math., 17 (1983), pp. 209-214.

[15] J. K. HALE AND C. ROCHA, *Bifurcation in a parabolic equation with variable diffusion*, Nonlinear Anal. Theory Methods Appl., 9 (1985), pp. 479-494.

[16] J. K. HALE AND K. SAKAMOTO, *Existence and stability of transition layers*, Japan J. Appl. Math., to appear.

[17] A. HASTINGS, *Spatial heterogeneity and the stability of predator-prey systems*, Theoret. Population Biol., 12 (1977), pp. 37-48.

[18] D. HENRY, *Geometric theory of semilinear parabolic equations*, Lecture Notes in Mathematics 840, Springer-Verlag, Berlin, New York, 1981.

[19] C. KELLET AND R. LUI, *Existence of steady-state solutions to predator-prey equations in a heterogeneous environment*, J. Math. Anal. Appl., 123 (1987), pp. 306-323.

[20] J. P. KEENER, *Causes of propagation failure in excitable media*, in Temporal Disorder in Human Oscillatory Systems, L. Rensing, U. Heiden, and M. MacKey, eds., Springer-Verlag, Berlin, New York, 1987, pp. 134-140.

[21] A. W. LEUNG AND B. BENDJILALI, *N species competition for a spatially heterogeneous prey with Neumann boundary conditions: Steady states and stability*, SIAM J. Appl. Math., 46 (1986), pp. 81-98.

[22] H. MATANO, *Asymptotic behavior and stability of solutions of semilinear diffusion equations*, J. Math. Kyoto Univ., 18 (1987), pp. 224-243.

[23] R. MCMURTRIE, *Persistence and stability of single species and prey-predator system in spatially heterogeneous environments*, Math. Biosci., 39 (1978), pp. 11-51.

[24] M. MIMURA AND Y. KAN-ON, *Predation-mediated coexistence and segregation structures*, in Pattern and Waves—Qualitative Analysis of Nonlinear Diffusion Equations, 1986, pp. 129-155.

[25] T. NAMBA, *Density-dependent dispersal and spatial distribution of a population*, J. Theoret. Biol., 86 (1980), pp. 351-363.

[26] T. NAMBA, *Competition for space in a heterogeneous environment*, preprint.
[27] A. OKUBO, *Ecology and Diffusion*, Lecture Notes in Biomath., Springer-Verlag, Berlin, New York, 1977.
[28] S. PACALA AND J. ROUGHGARDEN, *Spatial heterogeneity and interspecific competition*, Theoret. Population Biol., 21 (1982), pp. 92-113.
[29] C. V. PAO, *On nonlinear reaction-diffusion systems*, J. Math. Anal. Appl., 87 (1982), pp. 165-198.
[30] J. A. PAUWELUSSEN, *Nerve impulse propagation in a branching nerve system: a simple model*, Phys. D, 4 (1981), pp. 67-88.
[31] ———, *One way traffic of pulses in a neuron*, J. Math. Biol., 15 (1982), pp. 151-171.
[32] J. P. PAUWELUSSEN AND L. A. PELETIER, *Clines in the presence of asymmetric migration*, J. Math. Biol., 11 (1981), pp. 207-233.
[33] C. ROCHA, *Generic properties of equilibria of reaction diffusion equations with variable diffusion*, Proc. Roy. Soc. Edinburgh Sect. A, 101 (1985), pp. 45-55.
[34] N. SHIGESADA, *Spatial distribution of rapidly dispersing animals in heterogeneous environments*, Lecture Notes in Biomathematics 54, Springer-Verlag, Berlin, New York, 1984, pp. 478-491.
[35] N. SHIGESADA AND J. ROUGHGARDEN, *The role of rapid dispersal in the population dynamics of competition*, Theoret. Population Biol., 21 (1982), pp. 353-372.
[36] H. B. STEWART, *Generation of analytic semigroups by strongly elliptic operators under general boundary conditions*, Trans. Amer. Math. Soc., 259 (1980), pp. 299-310.
[37] S. YU, *Asymptotic behavior of solutions of heterogeneous reacting and diffusing systems*, Nonlinear Anal. Theory Meth. Appl., 9 (1985), pp. 275-288.
[38] E. YANAGIDA, *Stability of stationary distributions in space-dependent population growth process*, J. Math. Biol., 15 (1982), pp. 37-50.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 362-373, March 1990
© 1990 Society for Industrial and Applied Mathematics
005

THE STEFAN PROBLEM WITH A KINETIC CONDITION AT THE FREE BOUNDARY*

XIE WEIQING†

Abstract. This paper considers a class of one-dimensional solidification problems, in which a kinetic undercooling is incorporated into the temperature condition at the interface. A model problem with linear kinetic law is considered. This study indicates that the presence of a kinetic term at the interface can prevent finite-time blowup even though supercooling (superheating) exists. The mathematical effects of the kinetic term are discussed.

Key words. kinetic undercooling, Stefan problem, supercooled

AMS(MOS) subject classifications. 35K05, 35R35, 80A20

1. Introduction. Mathematical models of solidification that include interface kinetics effects have been considered for quite some time (see [1], [2], and references therein). This class of free boundary problems, which arise in a number of physical situations, is that of nonequilibrium problems, in which the phase-change temperature is dependent on the velocity of the front at which the phase change occurs (for more physical problems, see [3]-[6] and references therein). Here we study a model problem with linear kinetic law at the interface in the one-dimensional case. Specifically, let the curve $x=s(t)$ with $s(0)=b$ $(0<b<1)$ be defined as the interface that separates the liquid and solid phases. With u denoting temperature (scaled so that it vanishes at equilibrium), we may write the system of equations as

$$u_t = k_L u_{xx} \quad \text{in } Q_1=\{(x,t)/\ 0<x<s(t), 0<t\leqq T\}, \tag{1.1}$$

$$u_t = k_S u_{xx} \quad \text{in } Q_2=\{(x,t)/\ s(t)<x<1, 0<t\leqq T\}, \tag{1.2}$$

and on the interface $x=s(t)$ as

$$u^- = u^+ = u^I, \tag{1.3}$$

$$k_L u_x^- - k_S u_x^+ = -L\dot{s}(t), \tag{1.4}$$

$$u^I = \varepsilon \dot{s}(t), \tag{1.5}$$

$$s(0)=b, \qquad 0<b<1, \tag{1.6}$$

where k_L and k_S are thermal diffusivities of a liquid and a solid, respectively, $L\neq 0$ is the latent heat, ε is a constant, and the superscripts + and − denote, respectively, the right-hand and left-hand limits with respect to the spatial variable x. These equations are subject to the initial and boundary conditions

$$u(x,0)=\phi_1(x), \qquad 0\leqq x\leqq b, \tag{1.7}$$

$$u(x,0)=\phi_2(x), \qquad b\leqq x\leqq 1, \tag{1.8}$$

$$u(i-1,t)=f_i(t), \qquad t\geqq 0 \quad (i=1,2). \tag{1.9}$$

For the discussion below, we will also denote problem (1.1)-(1.9) as problem (P).

In the absence of the interface kinetics effects (i.e., when the coefficient $\varepsilon=0$), this problem is known as the Stefan problem, which has been widely studied and for which the mathematical results are fairly well understood.

* Received by the editors April 11, 1988; accepted for publication (in revised form) April 28, 1989. This work was supported by U.S. Navy grant N00014-86-G-0021.

† Mathematical Institute, 24/29 St. Giles, Oxford, OX1 3LB, United Kingdom.

The model in which the coefficient ε is nonzero has been considered by Coriell and Parker in [3], where the shape stability has been carried out for the effects of linear kinetic and of square kinetic law (i.e., $\varepsilon \dot{s} = u^2$) when temperature u satisfies Laplace's equation, in place of the time-dependent diffusion equation (see also the references in [3]). This has also been generalized by Coriell and Sekerka [4], who discuss the morphological stability of a planar solid-liquid interface during unidirectional solidification of a binary alloy. Crowley [5] has described several physical situations in which two kinds of nonequilibrium problems occur; those that arise in the modeling of alloy solidification in certain regimes, and those that arise in the study of condensed-phase flame propagation in which the reaction zone is thin. Visintin [7] has studied the latter problem using a variational approach, but has only been able to establish the existence of a generalized solution.

Our objective in this paper is to understand the mathematical effects of the kinetic term for the above model problem in the classical framework. The results here are completely parallel to the results that have been proved for the standard Stefan problem, but without sign restriction on the boundary and initial data that may lead to the finite-time blowup for the Stefan problem in certain circumstances. This indicates that the interface kinetics effect may regularize the problem, at least in the one-dimensional case, in such a way that it can stop blowup, even if supercooling (superheating) exists.

In § 2 we give the existence proof via a fixed-point argument. Section 3 establishes the uniqueness of the solution. Section 4 discusses the regularity of the free boundary (also for temperature $u(x, t)$), and the C^∞-regularity is proved by repeating the "bootstrap" process. Last, in § 5, we retrieve the solution of the Stefan problem by taking the limit $\varepsilon \to 0$, with sign restriction on the boundary and initial data and an additional assumption about the free boundary.

2. Existence of a solution of problem (P). In this section we establish the existence of problem (P), (1.1)-(1.9). A definition of a classical solution $(s(t); u(x, t))$ is defined in the usual sense, which satisfies

(i) $s \in C^1(0, T)$.

Denoting by $Q_T \equiv Q \equiv (0, 1) \times (0, T)$ and by u_i the restrictions to Q_i of $u(x, t)$,

(ii) $u_i(x, t) \in C(\bar{Q}_i) \cap C^{2,1}(Q_i)$,

$u_{ix} \in C(\bar{Q}_i \backslash \{x = i-1\})$, $\quad i = 1, 2,$

and (1.1)-(1.9), the functions $\phi_i(x)$ and $f_i(t)$ $(i = 1, 2)$ in (1.7)-(1.9) satisfy

$$f_i(t) \in C^1(\mathbb{R}^1) \cap L^\infty(\mathbb{R}^1), \quad \phi_1(x) \in C^1[0, b], \quad \phi_2(x) \in C^1[b, 1] \tag{2.1}$$

and the consistency conditions

$$f_1(0) = \phi_1(0), \quad f_2(0) = \phi_2(1), \quad \phi_1(b) = \phi_2(b). \tag{2.2}$$

To prove the existence result, we use a fixed-point argument.

Let $K(T_0, M) = \{s(t) \in C^1[0, T_0] / s(0) = b, 0 < s(t) < 1, |\dot{s}| \leqq M\}$, where M is a fixed constant to be specified below and T_0 is small enough so that

$$MT_0 \leqq \min\{b, 1-b\}. \tag{2.3}$$

For simplicity, we first consider the case in which the constants ε and L are positive.

For any given $s(t) \in \bar{K}(T_0, M)$, there exists a unique solution $u(x, t)$ of problem (1.1)-(1.3), (1.6)-(1.9), and

$$k_L u_x^-(s(t), t) - k_S u_x^+(s(t), t) + \frac{L}{\varepsilon} u(s(t), t) = 0 \tag{2.4}$$

(see, e.g., [8], [9]). With this choice of $u(x, t)$, we define the mapping F such that $Fs = h$ with

$$h(t) = b + \frac{1}{\varepsilon}\int_0^t u(s(\tau), \tau; s(t))\, d\tau, \tag{2.5}$$

where $u(x, t; s(t))$ is a solution of problems (1.1)-(1.3), (1.6)-(1.9), and (2.4) corresponding to the given interface $x = s(t)$.

If we can show that F has a fixed point that belongs in $K(T_0, M)$, then from (2.5) and (2.4), it follows that (1.4) and (1.5) are satisfied. Thus $(s(t), u(x, t))$ will then form a solution of problem (P).

To show that F maps $\bar{K}(T_0, M)$ into itself, we observe that $h(t)$ is again in $C^1[0, T]$ and $h(0) = b$ because $u(x, t)$ is a classical solution of (1.1)-(1.3), (1.6)-(1.9), and (2.4). Note that a straightforward application of maximum principle yields the estimate

$$\|u(x, t)\|_{0, Q_{T_0}} \leqq M_0 \equiv \max\,(\|\phi_i\|_0, \|f_i\|_0, i = 1, 2). \tag{2.6}$$

If we take $M = M_0/\varepsilon$, then by (2.5) we have $h(t) \in K(T_0, M)$. This means the mapping F is from $\bar{K}(T_0, M)$ into itself.

We next show that F is a continuous mapping. To do this, suppose $s^n(t), s(t) \in \bar{K}$ $(n = 1, 2, \cdots)$, $s^n \to s$ uniformly on $[0, T_0]$. Define $u^n(x, t)$, $u(x, t)$ to be the solution of problem (1.1)-(1.3), (1.6)-(1.9), (2.4) corresponding to the boundary $x = s^n(t)$, $x = s(t)$, respectively. We first prove that $u^n(x, t) \to u(x, t)$ in $C(\bar{Q}_{T_0})$.

Since the $u^n(x, t)$ is bounded uniformly, which is also the generalized solution of the problem

$$u_t - (K(x, t)u_x + a(x, t)u)_x + a(x, t)u_x = 0 \quad \text{in } Q_{T_0}, \tag{2.7}$$

with boundary and initial conditions (1.7)-(1.9) in the sense of [10], where

$$K(x, t) = \begin{cases} k_L, \\ k_S, \end{cases} \qquad a(x, t) = \begin{cases} 0 & \text{in } Q_1^n = \{0 < x \leqq s^n(t), 0 < t < T_0\}, \\ -L/\varepsilon & \text{in } Q_2^n = \{s^n(t) < x < 1, 0 < t < T_0\}, \end{cases} \tag{2.8}$$

then $u^n(x, t)$ has uniform Hölder constants on $\bar{Q}_{T_0}$ [10]. Thus there is a subsequence (which we also denote by $u^n(x, t)$) converging in $C(\bar{Q}_{T_0})$ to some function $u(x, t)$. Any region Q' bounded away from the interface $x = s(t)$ and parabolic boundary Γ of Q_{T_0} is also bounded away from $x = s^n(t)$ for n sufficiently large. The uniform bounds on higher derivatives of $u^n(x, t)$ in Q' allow us to pass to the limit to conclude that u satisfies (1.1), (1.2).

To prove (2.4), we note that the $u^n(x, t)$ satisfy

$$\int\int_{Q_{T_0}} [u_t^n \xi + K(x, t)u_x^n \xi_x + a(x, t)u^n \xi_x + a(x, t)u_x^n \xi]\, dx\, dt = 0$$

for all smooth $\xi(x, t)$ vanishing in a neighbourhood of Γ. Passing to the limit as $n \to \infty$ we have the same equality for u. As $\xi(x, t)$ is arbitrary and (1.1), (1.2), integrating by parts and noting that $s(t) \in \bar{K}(T_0, M)$ can show that (2.4) holds.

The uniqueness of the solution $u(x, t)$ of the boundary value problem (1.1)-(1.3), (1.6)-(1.9), (2.4) for given interface follows by standard methods (see, e.g., [8]). Then the limit function $u(x, t)$ of $u^n(x, t)$ is a solution of problem (1.1)-(1.3), (1.6)-(1.9), (2.4) corresponding to the interface $x = s(t)$; this implies that $u^n(x, t) \to u(x, t)$ in $C(\bar{Q}_{T_0})$.

We now prove $F(s^n) \to F(s)$ in $C[0, T_0]$, denoting by $\alpha(t) = \min(s^n(t), s(t))$ and by $u(x, t; \beta(t))$ the solution of (1.1)-(1.3), (1.6)-(1.9), (2.4) corresponding to the interface $x = \beta(t)$.

Then

$$
\begin{aligned}
F(s^n) - F(s) = & \frac{1}{\varepsilon} \int_0^t [u^n(s^n(\tau), \tau; s^n(t)) - u^n(\alpha(\tau), \tau; s^n(t))]\, d\tau \\
& + \frac{1}{\varepsilon} \int_0^t [u^n(\alpha(\tau), \tau; s^n(t)) - u(\alpha(\tau), \tau; s(t))]\, d\tau \\
& + \frac{1}{\varepsilon} \int_0^t [u(\alpha(\tau), \tau; s(t)) - u(s(\tau), \tau; s(t))]\, d\tau \\
\equiv & I_1^n + I_2^n + I_3^n .
\end{aligned} \tag{2.9}
$$

From the Hölder estimates of $u^n(x, t)$ and $u(x, t)$ on $\bar{Q}_{T_0}$ we get

$$
|I_i^n| \leqq \frac{M\alpha}{\varepsilon} \|s^n - s\|_{C[0, T_0]} \qquad (i = 1, 3), \tag{2.10}
$$

and the Dominated Convergence Theorem implies that

$$
|I_2^n| \to 0 \quad \text{as } n \to \infty \tag{2.11}
$$

because $u^n(x, t; s^n(t)) \to u(x, t; s(t))$ in $C(\bar{Q}_{T_0})$.

Hence, from (2.9)-(2.11), we get the continuity of the mapping F.

We have thus proved that the mapping F has a fixed point $\bar{K}(T_0, M)$ by the Schauder fixed-point theorem. The fixed point is also in $K(T_0, M)$. This gives a solution of problem (P) in $0 < t \leqq T_0$.

Noting that $u(x, T_0) \in C^1[0, s(T_0)] \cap C^1[s(T_0), 1]$ and (2.6), we can continue the process step by step to construct a global solution of problem (P) in $0 < t \leqq T$ by the same argument as that given above, where "global" means that either $T = \infty$ and $0 < s(t) < 1$ for $t < \infty$, or $T < +\infty$ and $\lim_{t \to T^-} s(t) = 0$ or 1. Thus the existence is proved for positive ε and L.

The method demonstrated above can clearly be extended to the case of negative ε and L.

In the case of $\varepsilon L < 0$, we note that the solution of (1.1)-(1.3), (1.6)-(1.9), (2.4) for given interface $s(t) \in \bar{K}(T_0, M)$ is also a generalized solution of problem (2.7), (1.7)-(1.9) in the sense of [10]. We then can obtain the L^∞-norm estimate of the solution $u(x, t)$, which may depend on $\|\phi_i\|_{L^\infty}$, $\|f_i\|_{L^\infty}$ $(i = 1, 2)$, ε, L, and T, but not on the interface $s(t)$ [10]. Using the same method as stated above we can prove the existence of a classical solution in $0 < t \leqq T_0$. We are careful to note also that the L^∞-norm estimate for $u(x, t)$ holds for any $T > 0$ for which (1.1)-(1.3), (1.6)-(1.9), (2.4) has a solution (and not just for the small T_0 for which existence is assured, as above). We then can extend the solution in $0 < t \leqq T_0$, step by step, to a global one; "global" here refers to existence either for all times $t \geqq 0$ or until the free boundary hits one of the fixed sides $x = 0$ or $x = 1$.

3. Uniqueness. We now prove the uniqueness of problem (1.1)-(1.9). Note that problem (1.1)-(1.9) corresponds to the Stefan problem if $\varepsilon = 0$; the uniqueness of the weak solution and hence of the classical solution has been proved by Oleinik [11]. Here we consider the uniqueness of solution of (1.1)-(1.9) when $\varepsilon \neq 0$.

We first derive some estimates for the solution of (1.1)-(1.9) for discussion below.

By a direct calculation of the integrals $\int_0^T d/dt \int_0^1 u^2\,dx\,dt$ and $\int_0^T d/dt \int_0^1 ku_x^2\,dx\,dt$ and using the boundary and interface conditions, we have

$$\begin{aligned}(3.1)\quad &\sup_{0\leqq t\leqq T}\int_0^1 u^2(x,t)\,dx+2\iint_{Q_1\cup Q_2} ku_x^2\,dx\,dt\\ &=\int_0^1 \phi^2(x)\,dx-\frac{2L}{\varepsilon}\int_0^T (u^I)^2\,dt-2\int_0^T (k_L f_1(t)u_x(0,t)-k_S f_2(t)u_x(1,t))\,dt,\end{aligned}$$

$$\begin{aligned}(3.2)\quad &\int_0^1 ku_x^2(x,T)\,dx+2\iint_{Q_1\cup Q_2} k^2u_{xx}^2\,dx\,dt\\ &=\int_0^1 k(\dot\phi(x))^2\,dx+2\int_0^T (k_L u_x^- u_t^- - k_S u_x^+ u_t^+)\,dt\\ &=\int_0^T (k_L(u_x^-)^2-k_S(u_x^+)^2)\dot s\,dt-2\int_0^T (k_L u_x(0,t)\dot f_1(t)-k_S u_x(1,t)\dot f_2(t))\,dt.\end{aligned}$$

Note that by (1.3)–(1.5) on $x=s(t)$, the second term on the right-hand side of (3.2) is equal to

$$-\frac{2L}{\varepsilon}\int_0^T u^I\frac{d}{dt}u^I-2\int_0^T (k_L(u_x^-)^2-k_S(u_x^+)^2)\dot s\,dt.$$

Then we get

$$\begin{aligned}(3.3)\quad \int_0^1 ku_x^2(x,T)\,dx+2\iint_{Q_1\cup Q_2} k^2u_{xx}^2\,dx\,dt &=\int_0^1 k(\dot\phi(x))^2\,dx-\frac{L}{\varepsilon}(u^2(s(T),T)-\phi^2(b))\\ &\quad-\int_0^T (k_L(u_x^-)^2-k_S(u_x^+)^2)\dot s\,dt\\ &\quad-2\int_0^T (k_L u_x(0,t)\dot f_1(t)-k_S u_x(1,t)\dot f_2(t))\,dt\end{aligned}$$

from (3.3), (3.4), and using the inequality below [12],

$$\|u_x(\cdot,T)\|^2_{L^\infty(a,b)}\leqq c\|u_{xx}\|_{L^2(a,b)}\|u_x\|_{L^2(a,b)},$$

we can derive

$$(3.4)\qquad \sup_{0\leqq t\leqq T}\int_0^1 ku_x^2(x,t)\,dx+\iint_{Q_1\cup Q_2}(ku_x^2+k^2u_{xx}^2)\,dx\,dt\leqq C,$$

where constant C depends on ε, L, T, $\|\phi_i\|_{H^1}$, $\|f_i\|_{H^1(0,T)}$ $(i=1,2)$, and $\delta=\min\{s(t),1-s(t),0\leqq t\leqq T\}$.

To prove the uniqueness, let $(s(t);u(x,t))$ and $(\underset{\sim}{s}(t);\underset{\sim}{u}(x,t))$ be two solutions in some time interval $0<t<T_1$, such that for some positive constant $\delta>0$,

$$\delta\leqq s(t),\qquad \underset{\sim}{s}(t)\leqq 1-\delta\quad\text{as } 0<t\leqq T_1,$$

we can prove that $s(t)=\underset{\sim}{s}(t)$, $u(x,t)=\underset{\sim}{u}(x,t)$ in $[0,T_1]$.

Define the new space variable ξ by

$$(3.5)\qquad \xi=\alpha(x,s(t)),$$

where the function $\alpha(x,s)$ as defined in $[0,1]\times[\delta,1-\delta]$ satisfies [13]

$$(3.6)\qquad \alpha(i,s)=i,\quad (i=0,1),\quad \alpha(s,s)=\tfrac12,\quad \alpha_x(s,s)=1\quad\text{as } \delta\leqq s\leqq 1-\delta,$$

$$(3.7)\qquad \alpha_x(x,s)\geqq\alpha_0>0,\quad |D^\beta\alpha|\leqq c\quad(|\beta|\leqq 3)\quad\text{in } [0,1]\times[\delta,1-\delta],$$

where α_0 and c are positive constants.

Then (3.5) determines x as a function of ξ and t, and the function $v(\xi, t)$, described by

$$v(\xi, t) = u(x, t), \tag{3.8}$$

is defined on $Q_T \equiv Q$ for any $0 < T < T_1$.

Setting

$$Q^- = (0, \tfrac{1}{2}) \times (0, T), \quad Q^+ = (\tfrac{1}{2}, 1) \times (0, T), \quad Q = Q^- \cup Q^+ \tag{3.9}$$

and noting that $u_x = v_\xi \alpha_x$, $u_{xx} = v_{\xi\xi}(\alpha_x)^2 + v_\xi \alpha_{xx}$, and $u_t = v_t + v_\xi \alpha_s \dot{s}(t)$, we find that the function $v(\xi, t)$ satisfies the system

$$v_t - k_L(\alpha_x)^2 v_{\xi\xi} = v_\xi(k_L \alpha_{xx} - \alpha_s \dot{s}(t)) \quad \text{in } Q^-, \tag{3.10}$$

$$v_t - k_S(\alpha_x)^2 v_{\xi\xi} = v_\xi(k_S \alpha_{xx} - \alpha_s \dot{s}(t)) \quad \text{in } Q^+, \tag{3.11}$$

$$v^-(\tfrac{1}{2}, t) = v^+(\tfrac{1}{2}, t) = \varepsilon \dot{s}(t), \qquad t > 0, \tag{3.12}$$

$$k_L v_\xi^-(\tfrac{1}{2}, t) - k_S v_\xi^+(\tfrac{1}{2}, t) = -L\dot{s}(t), \qquad t > 0, \tag{3.13}$$

and the initial and boundary conditions

$$v(\xi, 0) = \phi(\xi), \qquad 0 \leqq \xi \leqq 1, \tag{3.14}$$

$$v(i-1, t) = f_i(t), \quad i = 1, 2, \quad t \geqq 0, \tag{3.15}$$

where $\phi(\xi) = \phi_1$ if $\xi \in [0, \frac{1}{2}]$ and $\phi(\xi) = \phi_2$ if $\xi \in [\frac{1}{2}, 1]$; the superscripts + and − denote, respectively, the right-hand and left-hand limits with respect to the variable ξ.

Similarly, define $w(\xi, t) = \underset{\sim}{y}(x, t)$ with $\xi = \alpha(x, \underset{\sim}{s}(t))$. Then $w(\xi, t)$ satisfies a system similar to that satisfied by $v(\xi, t)$.

Now define

$$k = \begin{cases} k_L & \text{in } Q^-, \\ k_S & \text{in } Q^+, \end{cases} \tag{3.16}$$

$$z(\xi, t) = v(\xi, t) - w(\xi, t). \tag{3.17}$$

Then $z(\xi, t)$ satisfies the following system:

$$z_t - k(\alpha_x)^2 z_{\xi\xi} = g \quad \text{in } \underset{\sim}{Q}, \tag{3.18}$$

$$z^-(\tfrac{1}{2}, t) = z^+(\tfrac{1}{2}, t) = \varepsilon(\dot{s} - \dot{\underset{\sim}{s}}), \qquad t > 0, \tag{3.19}$$

$$k_L z_\xi^-(\tfrac{1}{2}, t) - k_S z_\xi^+(\tfrac{1}{2}, t) = -L(\dot{s} - \dot{\underset{\sim}{s}}), \qquad t > 0, \tag{3.20}$$

where

$$g = \begin{cases} g_1 & \text{in } Q^-, \\ g_2 & \text{in } Q^+, \end{cases} \tag{3.21}$$

$$g_1 = (k_L(\alpha_x)^2 - k_L(\underset{\sim}{\alpha}_x)^2) w_{\xi\xi} + (k_L \alpha_{xx} - \alpha_s \dot{s}) z_\xi + [(k_L \alpha_{xx} - \alpha_s \dot{s}) - (k_L \underset{\sim}{\alpha}_{xx} - \underset{\sim}{\alpha}_s \dot{\underset{\sim}{s}})] w_\xi, \tag{3.22}$$

$$g_2 = (k_S(\alpha_x)^2 - k_S(\underset{\sim}{\alpha}_x)^2) w_{\xi\xi} + (k_S \alpha_{xx} - \alpha_s \dot{s}) z_\xi + [(k_S \alpha_{xx} - \alpha_s \dot{s}) - (k_S \underset{\sim}{\alpha}_{xx} - \underset{\sim}{\alpha}_s \dot{\underset{\sim}{s}})] w_\xi \tag{3.23}$$

with $\underset{\sim}{\alpha} = \alpha(x, \underset{\sim}{s})$.

From (3.18) we get

$$\iint_{\underset{\sim}{Q}} z(z_t - k(\alpha_x)^2 z_{\xi\xi})\, d\xi\, dt = \iint_{\underset{\sim}{Q}} gz\, d\xi\, dt. \tag{3.24}$$

Note that

$$|g_i| \leqq c[|z_\xi| + |\dot{s} - \dot{\underset{\sim}{s}}| + |s - \underset{\sim}{s}| + |w_{\xi\xi}||s - \underset{\sim}{s}|] \qquad (i = 1, 2) \tag{3.25}$$

and the lower bound of the left-hand side of (3.24) is

$$\frac{1}{2}\int_0^1 z^2(\xi, T)\,d\xi + c\iint_{\underset{\sim}{Q}} (z_\xi)^2\,d\xi\,dt - c\iint_{\underset{\sim}{Q}} |z||z_\xi|\,d\xi\,dt, \tag{3.26}$$

while the right-hand side term of (3.24) is bounded by

$$\begin{aligned} &c\iint_{\underset{\sim}{Q}} (|z||z_\xi| + |\dot{s}-\dot{\underset{\sim}{s}}||z| + |s-\underset{\sim}{s}||z| + |w_{\xi\xi}||s-\underset{\sim}{s}||z|)\,d\xi\,dt \\ &\leqq c\left\{\int_0^T |\dot{s}-\dot{\underset{\sim}{s}}|^2\,dt + \eta\iint_{\underset{\sim}{Q}} ((z_\xi)^2 + (w_{\xi\xi})^2|s-\underset{\sim}{s}|^2)\,d\xi\,dt\right. \\ &\qquad \left. + c(\eta)\iint_{\underset{\sim}{Q}} z^2\,d\xi\,dt\right\}. \end{aligned} \tag{3.27}$$

The first term on the right-hand side of (3.27) is bounded by

$$\begin{aligned} c\int_0^T \frac{1}{\varepsilon^2} z^2\left(\frac{1}{2}, t\right) dt &\leqq \frac{c}{\varepsilon^2}\int_0^T \|z(\cdot, t)\|_{L^\infty}^2\,dt \\ &\leqq \eta\iint_{\underset{\sim}{Q}} (z_\xi)^2\,d\xi\,dt + c(\eta)\iint_{\underset{\sim}{Q}} z^2\,d\xi\,dt. \end{aligned} \tag{3.28}$$

Noting also that $\iint_{\underset{\sim}{Q}} |w_{\xi\xi}|^2\,d\xi\,dt \leqq c$ because of estimate (3.4), we have

$$\begin{aligned} \eta\iint_{\underset{\sim}{Q}} |w_{\xi\xi}|^2|s-\underset{\sim}{s}|^2\,d\xi\,dt &= \eta\int_0^T \left(\int_0^1 w_{\xi\xi}^2\,d\xi\right)|s-\underset{\sim}{s}|^2\,dt \\ &\leqq \frac{\eta}{\varepsilon}\int_0^T \left(\int_0^1 w_{\xi\xi}^2\,d\xi\right)\left(\int_0^t z^-\left(\frac{1}{2}, \tau\right)d\tau\right)^2 dt \\ &\leqq \frac{c\eta}{\varepsilon}\int_0^T \left(\int_0^1 w_{\xi\xi}^2\,d\xi\right)\int_0^t (\|z_\xi(\cdot, \tau)\|_{L^2(0,1/2)}^2 \\ &\quad + \|z(\cdot, \tau)\|_{L^2(0,1)}^2)\,d\tau\,dt \\ &\leqq \frac{c\eta}{\varepsilon}\iint_{\underset{\sim}{Q}} w_{\xi\xi}^2\,d\xi\,dt\left(\iint_{\underset{\sim}{Q}} (z^2 + z_\xi^2)\,d\xi\,dt\right). \end{aligned} \tag{3.29}$$

From (3.26)–(3.29) and choosing η sufficiently small, we get

$$\frac{1}{2}\int_0^1 z^2(\xi, \tau)\,d\xi + c\iint_{\underset{\sim}{Q}} z_\xi^2\,d\xi\,dt \leqq c\int_0^T\int_0^1 z^2(\xi, t)\,d\xi\,dt.$$

This implies that

$$\int_0^1 z^2(\xi, \tau)\,d\xi = 0$$

so that $z \equiv 0$ on Q and consequently $v(\xi, t) = w(\xi, t)$, $s(t) = \underset{\sim}{s}(t)$. Thus the uniqueness is proved.

4. Regularity of the solution. We now discuss the regularity of the solution for problem (P). Recall that, for the ordinary Stefan problem, the free boundary is always infinitely differentiable on the time interval in which existence is assured [14], [15]. Here the same result is proved for the problem with linear interface kinetics (1.5).

To prove the regularity of the free boundary, we consider a neighbourhood of the free boundary $x=s(t)$ for $0<t<T$, where $[0, T]$ is the time interval in which the solution of problem (P) exists with $\min\{s(t), 1-s(t), 0\leqq t\leqq T\}>0$.

Choose $\delta>0$ so small that the region

$$N=\{(x, t)/\ s(t)-\delta<x<s(t)+\delta, 0<t<T\}$$

lies in Q_T. Change the variable to

$$\xi=x-s(t) \tag{4.1}$$

and let $v(\xi, t)=u(\xi+s(t), t)$ for $(\xi, t)\in N_1$, where $(s(t), u(x, t))$ is the unique solution of problem (P), $N_1=\{(\xi, t)/-\delta<\xi<\delta, 0<t<T\}$.

Then $v(\xi, t)$ satisfies the following system:

$$v_t-k_{\mathrm{L}}v_{\xi\xi}=\dot{s}(t)v_\xi, \quad -\delta<\xi<0, \quad 0<t<T, \tag{4.2}$$

$$v_t-k_{\mathrm{S}}v_{\xi\xi}=\dot{s}(t)v_\xi, \quad 0<\xi<\delta, \quad 0<t<T, \tag{4.3}$$

$$v(\xi, 0)=\underset{\sim}{\phi}(\xi), \quad -\delta\leqq\xi\leqq\delta, \tag{4.4}$$

$$v^-(0, t)=v^+(0, t)=\varepsilon\dot{s}(t), \quad 0<t<T, \tag{4.5}$$

$$k_{\mathrm{L}}v_\xi^-(0, t)-k_{\mathrm{S}}v_\xi^+(0, t)=-\dot{s}(t) \quad 0<t<T. \tag{4.6}$$

We assume, without loss of generality, that $k_{\mathrm{L}}\geqq k_{\mathrm{S}}$ and define the following two functions:

$$w(\xi, t)=v(\xi, t)-v(-\underset{\sim}{k}\xi, t), \tag{4.7}$$

$$\underset{\sim}{w}(\xi, t)=v(\xi, t)+\underset{\sim}{k}v(-\underset{\sim}{k}\xi, t), \tag{4.8}$$

where

$$\underset{\sim}{k}=(k_{\mathrm{S}}/k_{\mathrm{L}})^{1/2}\leqq 1. \tag{4.9}$$

Then $w(\xi, t)$ satisfies

$$w_t-k_{\mathrm{L}}w_{\xi\xi}=\dot{s}(t)(v_\xi(\xi, t)-v_\xi(-\underset{\sim}{k}\xi, t)) \quad \text{in } N_1^-, \tag{4.10}$$

$$w(0, t)=0, \quad 0\leqq t\leqq T, \tag{4.11}$$

$$w(\xi, 0)=\underset{\sim}{\phi}_1(\xi)-\underset{\sim}{\phi}_2(-\underset{\sim}{k}\xi), \quad -\delta\leqq\xi\leqq 0, \tag{4.12}$$

and $\underset{\sim}{w}(\xi, t)$ satisfies

$$\underset{\sim}{w}_t-k_{\mathrm{L}}\underset{\sim}{w}_{\xi\xi}=\dot{s}(t)(v_\xi(\xi, t)+\underset{\sim}{k}v_\xi(-\underset{\sim}{k}\xi, t)) \quad \text{in } N_1^-, \tag{4.13}$$

$$k_{\mathrm{L}}\left(\underset{\sim}{w}_\xi(0, t)+\frac{1}{(\underset{\sim}{k}+1)\varepsilon}\underset{\sim}{w}(0, t)\right)=0, \quad 0\leqq t\leqq T, \tag{4.14}$$

$$\underset{\sim}{w}(\xi, 0)=\underset{\sim}{\phi}_1(\xi)+\underset{\sim}{k}\underset{\sim}{\phi}_2(-\underset{\sim}{k}\xi), \quad -\delta\leqq\xi\leqq 0, \tag{4.15}$$

where $N_1^-=\{(\xi, t)/-\delta<\xi<0, 0<t<T\}$. Note that the functions on the right-hand side of (4.10), (4.13) belong to L^∞. According to the local parabolic L^p-estimates for the problem with the Dirichlet boundary condition and the problem with directional derivative [10], we have

$$\|w, \underset{\sim}{w}\|_{W_p^{2,1}(N_2)}\leqq C, \tag{4.16}$$

where $p>1$ is arbitrarily constant, for any interior domain N_2, the boundary of which contains a segment $\xi=0$. We choose $p>1$ suitable large such that

$$w_\xi,\ \underset{\sim}{w}_\xi \in C^{\alpha,\alpha/2}(\bar{N}_2), \tag{4.17}$$

where $0<\alpha<1$ [10].

Note that, from (4.7) and (4.8),

$$v(\xi,t)=\frac{1}{\underset{\sim}{k}+1}(\underset{\sim}{k}w(\xi,t)+\underset{\sim}{w}(\xi,t)), \tag{4.18}$$

$$v(-\underset{\sim}{k}\xi,t)=\frac{1}{\underset{\sim}{k}+1}(\underset{\sim}{w}(\xi,t)-w(\xi,t)), \tag{4.19}$$

and hence $v_\xi(\xi,t)\in C^{\alpha,\alpha/2}(\overline{N_3^-})\cap C^{\alpha,\alpha/2}(\overline{N_3^+})$, where N_3^- is a domain of the form $(-\delta,0)\times(\eta,T)$ and N_3^+ is its reflection in the line $\xi=0$. By (4.6) we have $\dot{s}(t)\in C^{\alpha/2}(0,T)$, and so the functions on the right-hand sides of (4.10), (4.13) are now $C^{\alpha,\alpha/2}(\overline{N_3^-})$. Then, according to the parabolic Schauder estimates [10], [16] for (4.10), (4.13), we have

$$w(\xi,t),\ \underset{\sim}{w}(\xi,t)\in C^{2+\alpha,(2+\alpha)/2}(\overline{N_4^-}),$$

where N_4^- is an interior domain of N_3^-, the boundary of which contains a segment of the $\xi=0$ axis; and so $w_\xi(\xi,t)$, $\underset{\sim}{w}_\xi(\xi,t)\in C^{1+\alpha,(1+\alpha)/2}(\overline{N_4^-})$. Using (4.18), (4.19), and (4.6), we find that the functions on the right-hand side of (4.10), (4.13) are $C^{1+\alpha,(1+\alpha)/2}(\overline{N_4^-})$.

This "bootstrap" process may now be continually repeated, each time to derive better estimates on the derivatives of w, $\underset{\sim}{w}$ and therefore of $s(t)$, $V(\xi,t)$, $v(-\underset{\sim}{k}\xi,t)$, all the way up to the $\xi=0$ axis for any $t>0$. Hence $s(t)$ is infinitely differentiable in the time interval $0<t<T$.

By standard parabolic regularity theory [12], [17], we can also obtain that $u(x,t)$ is infinitely differentiable in $Q_1\cup\{x=s(t),\ t>0\}$ and $Q_2\cup\{x=s(t),\ t>0\}$.

5. The limit process as $\varepsilon\to 0$. We now discuss the limit process for ε when sign restrictions are imposed on ε, L, and boundary and initial data. This means we will henceforth assume that

$$f_1(t)>0,\quad f_2(t)<0,\quad \phi_1(x)\geqq 0,\quad \phi_2(x)\leqq 0, \tag{5.1}$$

where $\phi_1(b)=\phi_2(b)=0$ and ε, L are positive constants.

We will retrieve the solution of the standard Stefan problem by taking the limit $\varepsilon\to 0^+$ in the classical sense, and this only with an additional restriction about monotonicity (nondecreasing) of the free boundary.

We recall from previous sections that problem (1.1)-(1.9) possesses a unique classical solution and note that the solution $(s_\varepsilon(t);\ u_\varepsilon(x,t))$ $(\varepsilon>0)$ satisfies

$$u_\varepsilon(s_\varepsilon(t),t)=\varepsilon\dot{s}_\varepsilon(t), \tag{5.2}$$

$$Ls_\varepsilon(t)=L_b+\int_0^1\phi(x)\,dx-\int_0^1 u_\varepsilon(x,t)\,dx-\int_0^t(k_L u_{\varepsilon_x}(0,\tau)-k_S u_{\varepsilon_x}(1,\tau))\,d\tau \tag{5.3}$$

and estimates (2.6). To pass the limit $\varepsilon\to 0$ in the classical sense, we need to derive some a priori estimates for the solution $(s_\varepsilon(t);\ u_\varepsilon(x,t))$, which is independent of ε.

Using the method employed in [8], we can derive an estimate of $(1/\varepsilon)u_\varepsilon(s_\varepsilon(t),t)$ that is independent of ε. This is done by comparing the function $v(x,t)$, which satisfies (1.1)-(1.3), (1.7)-(1.9), and the interface condition $v=0$. Note carefully that

$v_x^{\pm}(s_\varepsilon(t), t) < 0$ by the maximum principle and that the lower bound of $v_x^-(s_\varepsilon(t), t)$ can be estimated by using the auxiliary function method that is independent of ε [9]; this is derived from the crucial assumption about nondecreasing of the free boundary.

Consider the function $(1/\varepsilon)(u_\varepsilon - v)$ and suppose that p_0 is a positive maximum point of that function. Then p_0 must belong to $\{(x, t)/\ x = s_\varepsilon(t), 0 < t \leqq T\}$, $(u_{\varepsilon_x}^- - v_x^-)(p_0) > 0$, and $(u_{\varepsilon_x}^+ - v_x^+)(p_0) < 0$ by the strong maximum principle. Furthermore, we have

$$\begin{aligned} \max_{Q_T} \frac{L}{\varepsilon}(u_\varepsilon - v) &= \frac{L}{\varepsilon}(u_\varepsilon - v)(p_0) = \frac{L}{\varepsilon} u_\varepsilon(p_0) \\ &= -k_L(u_{\varepsilon_x}^- - v_x^-)(p_0) + k_S(u_{\varepsilon_x}^+ - v_x^+)(p_0) - k_L v_x^-(p_0) + k_S v_x^+(p_0) \\ &\leqq -k_L v_x^-(p_0) \leqq B \quad \text{(independent of } \varepsilon); \end{aligned} \tag{5.4}$$

this implies that

$$|\dot{s}_\varepsilon(t)| \leqq C, \tag{5.5}$$

where C is a constant that is independent of ε. From (3.1) and (3.3) and using (5.5), we have

$$\begin{aligned} &\sup_{0 \leqq t \leqq T} \int_0^1 u_\varepsilon^2(x, t)\, dx + \int_0^1 k u_{\varepsilon_x}^2(x, T)\, dx + 2 \iint_{Q_1 \cup Q_2} k^2 u_{\varepsilon_{xx}}^2\, dx\, dt \\ &\leqq C \Big[\|\phi\|_{H^1(0,1)}^2 + \|f_1\|_{H^1(0,T)}^2 + \|f_2\|_{H^1(0,T)}^2 \\ &\qquad + \int_0^T (\|u_{\varepsilon_x}(\cdot, t)\|_{L^2(0, s_\varepsilon(t))}^2 + \|u_{\varepsilon_x}(\cdot, t)\|_{L^2(s_\varepsilon(t), 1)}^2)\, dt \Big] \\ &\leqq C \Big[\|\phi\|_{H^1(0,1)}^2 + \|f_1\|_{H^1(0,T)}^2 + \|f_2\|_{H^1(0,T)}^2 + \eta \iint_{Q_1 \cup Q_2} u_{\varepsilon_{xx}}^2\, dx\, dt \\ &\qquad + c(\eta) \int_0^T \int_0^1 u_{\varepsilon_x}^2\, dx\, dt \Big]. \end{aligned} \tag{5.6}$$

Then we derive

$$\begin{aligned} &\sup_{0 \leqq t \leqq T} \int_0^1 (u_\varepsilon^2(x, t) + u_{\varepsilon_x}^2(x, t))\, dx + \iint_{Q_1 \cup Q_2} (u_{\varepsilon_x}^2 + u_{\varepsilon_{xx}}^2)\, dx\, dt \\ &\qquad \leqq C(T, \|\phi\|_{H^1(0,1)}^2 + \|f_i\|_{H^1(0,T)}^2 (i = 1, 2)). \end{aligned} \tag{5.7}$$

We can then get the solution of the Stefan problem via compactness arguments, and possibly take subsequences. Indeed, in the light of estimates (5.5) and (5.7), there exist a couple $(s(t); u(x, t))$ with $s(t) \in C^{0,1}[0, T]$ and $u \in C^{\alpha, \alpha/2}(\bar{Q}_T)$ $(0 < \alpha < 1)$, such that $s_\varepsilon \to s$ uniformly in $C[0, T]$ and $u_\varepsilon \to u$ uniformly in $C(\bar{Q}_T)$; moreover, $(s(t); u(x, t))$ satisfying (1.1), (1.2), (1.7)-(1.9), and (5.2), (5.3) imply, respectively,

$$u = 0 \quad \text{on } x = s(t), \tag{5.8}$$

$$Ls(t) = b + \int_0^1 \phi(x)\, dx - \int_0^t u(x, t)\, dx + \int_0^t (k_L u_x(0, \tau) - k_S u_x(1, \tau))\, d\tau. \tag{5.9}$$

Note that $s(t) \in C^{0,1}[0, T]$ and then $u_x \in C(\bar{Q}_1) \cap C(\bar{Q}_2)$; then (5.8) implies the Stefan condition (1.4).

Our demonstration above also indicates that we can retrieve the solution of the one-phase Stefan problem by taking $\varepsilon \to 0^+$ without any additional restrictions (the

monotonicity of the free boundary, in fact, holds automatically in this case). Note that from (3.1), (3.3), we can get

$$\sup_{0\leqq t\leqq T}\int_0^{s_\varepsilon(t)} u_{\varepsilon_x}^2(x,t)\,dx+\int_0^T\int_0^{s_\varepsilon(t)}(u_{\varepsilon_x}^2+u_{\varepsilon_{xx}}^2)\,dx\,dt\leqq C, \tag{5.10}$$

where the constant C depends only on the known data, but not on ε.

By the interpolation inequality [12], $\|f\|^2_{L^\infty(a,b)}\leqq C\|f_x\|_{L^2(a,b)}\|f\|_{L^2(a,b)}$, (5.10) implies that

$$\int_0^T \|u_{\varepsilon_x}(\cdot,t)\|^4_{L^\infty(0,s_\varepsilon(t))}\,dt\leqq C,$$

and then $\|\dot{s}_\varepsilon(t)\|_{L^4(0,T)}\leqq C$, where the constant C is independent of ε. With the above estimates at hand, we can easily get the solution of the one-phase Stefan problem via compactness arguments.

It is perhaps also worth noting that, in general, without the restriction (5.1), it may not always be expected to pass the limit $\varepsilon\to 0$ because the finite-time blowup might, in fact, occur for the supercooled (superheated) Stefan problem in certain circumstances (see, e.g., [18]-[20]).

Acknowledgments. I thank Dr. J. R. Ockendon for his helpful advice and encouraging discussions during the preparation of this paper. I am also very grateful to the referees for suggesting several improvements and pointing out misprints in the original version of the manuscript.

REFERENCES

[1] J. W. CAHN, W. B. HILLIG, AND G. W. SEARS, *The molecular mechanism of solidification*, Acta Mech., 12 (1964), pp. 1421-1439.

[2] B. CHALMERS, *Principles of Solidification*, Krieger, New York, 1977.

[3] S. R. CORIELL AND R. L. PARKER, *Interface kinetics and the stability of the shape of a solid sphere growing from the melt*, in Proc. International Conference on Crystal Growth, Boston, 1966, pp. 20-24.

[4] S. R. CORIELL AND R. F. SEKERKA, *Oscillatory morphological instabilities due to non-equilibrium segregation*, J. Crystal Growth, 61 (1983), pp. 499-508.

[5] A. B. CROWLEY, *Some remarks on a non-equilibrium solidification problem*, in Free and Moving Boundary Problems, K. H. Hoffmann and J. Sprekels, eds., Pitman, Boston, 1989.

[6] G. CAGINALP, *A conserved phase field system; Implications for kinetics undercooling*, Phys. Rev. B, to appear.

[7] A. VISINTIN, *Stefan problem with a kinetic condition at the free boundary*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 97-122.

[8] L. JIANG AND W. XIE, *A parabolic equation with discontinuous coefficients and a Signorini-type interface condition*, Acta Sci. Natur. Univ. Pekinensis, 3 (1986), pp. 1-14.

[9] ———, *A two-phase Stefan-Signorini problem*, Acta Sci. Natur. Univ. Pekinensis, 5 (1986), pp. 1-14.

[10] O. A. LADYZHENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and quasi-linear equations of parabolic type*, AMS Transl. 23, American Mathematical Society, Providence, RI, 1968.

[11] O. A. OLEINIK, *A method of solution of the general Stefan problem*, Soviet Math., 1 (1960), pp. 1350-1354.

[12] A. FRIEDMAN, *Partial Differential Equations*, Holt, Reinhart, and Winston, New York, 1969.

[13] C. BAIOCCHI, L. C. EVANS, L. FRANK, AND A. FRIEDMAN, *A uniqueness for two immiscible fluids in a one dimensional porous medium*, J. Differential Equations, 36 (1980), pp. 249-256.

[14] L. JIANG, *Existence and differentiability of the solution of the two-phase Stefan problem for quasilinear parabolic equations*, Acta Math. Sinica, 15 (1965), pp. 749-764.

[15] D. SCHAEFFER, *A new proof of the infinite differentiability of the free boundary in the Stefan problem*, J. Differential Equations, 20 (1976), pp. 266-269.

[16] L. I. KAMYNIN AND V. N. MASLENNIKOVA, *Boundary estimates of the solution of the third boundary problem for a parabolic equation*, Dokl. Akad. Nauk SSSR, 153 (1963), pp. 526-529.

[17] M. GEVREY, *Sur les équations aux dérivées partielles du type parabolique*, J. Math. Pures Appl., 9 (1913), pp. 305–475.

[18] A. FASANO, M. PRIMICERIO, AND A. A. LACEY, *New results on some classical parabolic free-boundary problems*, Quart. Appl. Math., 38 (1981), pp. 439–460.

[19] A. A. LACEY AND J. R. OCKENDON, *Ill-posed boundary problems*, Control Cybernet., 14 (1985), pp. 275–296.

[20] S. D. HOWISON, J. R. OCKENDON, AND A. A. LACEY, *Singularity development in moving boundary problems*, Quart. J. Mech. Appl. Math., 38 (1985), pp. 343–360.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 374–393, March 1990
©1990 Society for Industrial and Applied Mathematics
006

SEMIGROUP THEORY AND NUMERICAL APPROXIMATION FOR EQUATIONS IN LINEAR VISCOELASTICITY*

R. H. FABIANO† AND K. ITO‡

Abstract. The following abstract integro-differential equation

$$\ddot{u}(t) + A\left[Eu(t) - \int_{-r}^{0} g(s)u(t+s)\,ds\right] = f(t)$$

is considered on a Hilbert space. Such equations arise in the modeling of linear viscoelastic beams. The equation is reformulated as an abstract Cauchy problem, and several approximation schemes are discussed. Well-posedness and convergence results are given in the context of linear semigroup theory. Results of numerical eigenvalue calculations for various approximation schemes are discussed.

Key words. integro–differential equation, linear semigroup theory

AMS(MOS) subject classifications. 45K05, 65M60, 73F99

1. Introduction.

We consider the following integro-differential equation:

$$\ddot{u}(t) + A\left[Eu(t) - \int_{-r}^{0} g(s)u(t+s)\,ds\right] = f(t). \tag{1.1}$$

Here A is a positive definite, self-adjoint unbounded operator on a Hilbert space H, and $f(t)$ is an H-valued function. Also, E is a positive constant (a stiffness coefficient in applications to linear viscoelasticity) and $g(s)$ is a "history kernel" which is further characterized below. In §2, we formulate the equation as an abstract Cauchy problem and prove well-posedness. This abstract framework is useful for the construction and convergence analysis of approximation schemes for (1.1). It can be seen that an approximation scheme for this type of equation involves discretization of the spatial variable (finite elements, for example) together with an appropriate approximation scheme for the resulting delay equation (we consider the so-called averaging scheme of Banks and Burns [1] and a newer scheme recently developed by Ito and Kappel [10].) In §3, we discuss this approach to approximation, and prove a convergence result using the Trotter–Kato semigroup approximation theorem.

The results here extend those of [2]. Our theory allows for a singularity at the origin in the kernel $g(s)$, while in [2] it is required that $g(s)$ be bounded. Less importantly, there is a possibly restrictive technical condition in the convergence proof of [2] which we do not require.

One of our motivations for studying (1.1) is that this type of equation arises in the modeling of linear viscoelastic beams (see [4]–[9]). The abstract framework that

*Received by the editors June 29, 1988; accepted for publication (in revised form) March 14, 1989. This research was carried out while the authors were members of the Center for Control Sciences, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The research of both authors was supported in part by the Air Force Office of Scientific Research under grant AFOSR–84–0398 and contract AFOSR–F49620–86–C–0111, the National Science Foundation under grant MCS–8504316, and the National Aeronautics and Space Administration under NASA grant NAG–1–517.

†Department of Mathematics, Texas A&M University, College Station, Texas 77843.

‡Center for Applied Mathematical Sciences, DRB 306, University of Southern California, Los Angeles, California 90089-1113.

we consider is especially useful for the applications to parameter identification and control problems for viscoelastic models. (We do not discuss these applications here; see [2], [3]). It is also of interest to investigate the damping induced by the assumption of Boltzmann viscoelasticity in modeling the motion of a beam. This is related to the behavior of the eigenvalues. In the purely elastic model (no damping) the eigenvalues of the underlying dynamical system lie on the imaginary axis. A good indication of the damping behavior in the viscoelastic model is the location of the eigenvalues in the left halfplane. In §4, we calculate these eigenvalues, and we also discuss the convergence behavior of the approximation schemes developed in §3. In addition, we consider a new approximation scheme based on a variation of the averaging scheme in which a nonuniform mesh is used. The theoretical convergence arguments for this scheme, which are similar to those used for the averaging scheme, are not included here. However, we do include some numerical examples for this scheme to illustrate the fast convergence which we have observed.

2. An abstract Cauchy problem. In this section we reformulate (1.1) as an abstract Cauchy problem and prove well-posedness. First we establish some preliminary ideas. Assume V and H are Hilbert spaces and $V \subset H$ with continuous dense injection. Let V^* denote the strong dual space of V. We identify H with its dual, so that $V \subset H = H^* \subset V^*$. The dual product $\langle \phi, \psi \rangle_{V,V^*}$ on $V \times V^*$ is the unique extension by the continuity of the scalar product $\langle \phi, \psi \rangle_H$ of H restricted to $V \times H$. Consider a symmetric sesquilinear form σ on V such that

$$|\sigma(u,v)| \leq C|u|_V\,|v|_V \quad \text{for } u,v \in V \tag{2.1}$$

$$\sigma(u,u) \geq \omega |u|_V^2 \quad \text{for } u \in V \tag{2.2}$$

where $\omega > 0$. Let $A \in \mathcal{L}(V, V^*)$ be defined by

$$\sigma(u,v) = \langle Au, v \rangle_{V^*,V} \quad \text{for all } u,v \in V. \tag{2.3}$$

It then follows from [16, Thm. 2.2.3] that the restriction of A on H with

$$\operatorname{dom}(A) = \{u \in V : Au \in H\}, \tag{2.4}$$

where we will use the same symbol A for such a restriction, defines a positive definite and self-adjoint operator on H, $\operatorname{dom}(A^{1/2}) = V$, and

$$\sigma(u,v) = \langle A^{1/2}u, A^{1/2}v \rangle \quad \text{for } u,v \in V.$$

Thus, V can be equipped with the scalar product $\langle u, v \rangle_V = \sigma(u, v,)$.

Given $r \in [-\infty, 0)$ consider the second-order equation in H

$$\ddot{u}(t) + A\left(Eu(t) - \int_{-r}^{0} g(\theta)u(t+\theta)\,d\theta\right) = f(t) \tag{2.5}$$

where E is a positive constant, $t \to f(t)$ is an H-valued integrable function, and we assume $g > 0$ and $g' \geq 0$ on $(-r, 0)$, and g is integrable with

$$\alpha = E - \int_{-r}^{0} g(\theta)\,d\theta > 0. \tag{2.6}$$

An example is

$$g(\theta) = \frac{ce^{\gamma\theta}}{|\theta|^p} \tag{2.7}$$

where $c > 0$ and $0 \leq p < 1$ and $\gamma > 0$.

Next, let $W = L_g^2(-r, 0; V)$ be the Hilbert space of all V-valued, square integrable functions defined on the measure space $([-r, 0], \text{Borel sets}, g\,d\theta)$, equipped with norm

$$|w|_W^2 = \int_{-r}^{0} g(\theta)|w(\theta)|_V^2\,d\theta. \tag{2.8}$$

Let Z denote the Hilbert space $V \times H \times W$ equipped with norm

$$|(u, v, w)|_Z^2 = \alpha|u|_V^2 + |v|_H^2 + \int_{-r}^{0} g(\theta)|w(\theta)|_V^2\,d\theta. \tag{2.9}$$

Motivated by the state-space formulations found in [6] and [17], we define for $t \geq 0$

$$\begin{aligned} v(t) &= \dot{u}(t) \\ w(t, \theta) &= u(t) - u(t + \theta), \qquad -r \leq \theta \leq 0. \end{aligned} \tag{2.10}$$

Then $z = (u, v, w) \in Z$ formally satisfies

$$\frac{d}{dt}z(t) = \mathcal{A}z(t) + \text{col}[0, f(t), 0] \tag{2.11}$$

where

$$\mathcal{A}z = \left(v, -A\left(\alpha u + \int_{-r}^{0} g(\theta)w(\theta)\,d\theta\right), v + \frac{d}{d\theta}w\right) \tag{2.12}$$

for $z = (u, v, w) \in Z$.

Thus (1.1) has been reformulated as the abstract Cauchy problem (2.11). Next we use the Lumer–Phillips theorem [13] to show that $\mathcal{A}$ generates a C_0-semigroup on Z.

THEOREM 2.1. *The linear operator $\mathcal{A}$ on Z, defined by* (2.12) *with domain*

$$\text{dom}(\mathcal{A}) = \left\{ \begin{array}{l} (u, v, w) \in Z : v \in V \\ \dot{w} \in L_g^2(-r, 0; V), \quad w(0) = 0, \\ \alpha u + \int_{-r}^{0} g(\theta)w(\theta)\,d\theta \in \text{dom}(A) \end{array} \right\} \tag{2.13}$$

generates a C_0-semigroup $S(t)$ on Z.

Proof. First we argue the dissipativeness. For $z = (u, v, w) \in \text{dom}(\mathcal{A})$

$$\begin{aligned} \langle \mathcal{A}z, z\rangle &= \alpha\langle u, v\rangle_V - \left\langle A\left(\alpha u + \int_{-r}^{0} g(\theta)w(\theta)\,d\theta\right), v\right\rangle_{V^*, V} \\ &\quad + \int_{-r}^{0} g(\theta)\langle v + Dw, w\rangle_V\,d\theta \\ &= \int_{-r}^{0} g(\theta)\langle Dw, w\rangle_V\,d\theta \end{aligned}$$

where

$$D\phi = \frac{d}{d\theta}\phi \quad \text{with } \operatorname{dom}(D) = \{\phi \in W : D\phi \in W \text{ and } \phi(0) = 0\} \tag{2.14}$$

and we used $\langle u, v\rangle_V = \sigma(u, v) = \langle Au, v\rangle_{V^*,V}$ for $u, v \in V$. Let us consider for $\epsilon > 0$ and $R < r$,

$$\begin{aligned} I_{\epsilon,R} &= \int_{-R}^{-\epsilon} \frac{1}{2} g(\theta) \frac{d}{d\theta} |w(\theta)|_V^2 \, d\theta \\ &= \frac{1}{2} g(-\epsilon)|w(-\epsilon)|_V^2 - \frac{1}{2} g(-R)|w(-R)|_V^2 - \frac{1}{2}\int_{-R}^{-\epsilon} g'(\theta)|w(\theta)|_V^2 \, d\theta \\ &\le \frac{1}{2} g(-\epsilon)|w(-\epsilon)|_V^2 . \end{aligned} \tag{2.15}$$

Since $w(-\epsilon) = w(0) - \int_{-\epsilon}^0 Dw(\theta)\, d\theta = -\int_{-\epsilon}^0 Dw(\theta)\, d\theta$, by the Cauchy–Schwarz inequality

$$|w(-\epsilon)|_V^2 \le \int_{-\epsilon}^0 \frac{d\theta}{g(\theta)} \int_{-\epsilon}^0 g(\theta)|Dw|_V^2 \, d\theta.$$

Note that $g(-\epsilon)\int_{-\epsilon}^0 d\theta/g(\theta) = \int_{-\epsilon}^0 g(-\epsilon)/g(\theta)\, d\theta \le \epsilon$. Thus, we obtain

$$I_{\epsilon,R} \le \epsilon \int_{-\epsilon}^0 g(\theta)|Dw|_V^2 \, d\theta,$$

and this implies for $z \in \operatorname{dom}(\mathcal{A})$ that

$$\langle \mathcal{A}z, z\rangle = \lim_{\epsilon \downarrow 0, R \uparrow r} I_{\epsilon,R} \le 0. \tag{2.16}$$

Next we will show that $(\lambda I - \mathcal{A})\operatorname{dom}(\mathcal{A}) = Z$ for $\operatorname{Re}\lambda > 0$. The equation $(\lambda I - \mathcal{A})z = (\phi, \psi, h)$ is written as

$$\lambda u - v = \phi \in V \tag{2.17}$$

$$\lambda v + A\left(\alpha u + \int_{-r}^0 g(\theta) w(\theta)\, d\theta\right) = \psi \in H \tag{2.18}$$

$$\lambda w - v - Dw = h \quad \text{and} \quad w(0) = 0. \tag{2.19}$$

From (2.19), we obtain

$$w(\theta) = \int_\theta^0 e^{\lambda(\theta-\xi)}\,(v + h(\xi))\, d\xi .$$

From (2.17), $v = \lambda u - \phi$. Substituting these into (2.18), we have

$$\lambda^2 u + A\left[\alpha u + \int_{-r}^0 g(\theta) \int_\theta^0 e^{\lambda(\theta-\xi)}\,(\lambda u - \phi + h(\xi))\, d\xi\, d\theta\right] = \psi + \lambda\phi,$$

or

$$\begin{aligned} &\left[\lambda^2 + A\left(E - \int_{-r}^0 e^{\lambda\theta} g(\theta)\, d\theta\right)\right] u \\ &\qquad = \psi + \lambda\phi + \int_{-r}^0 e^{\lambda\theta} g(\theta) \int_\theta^0 e^{-\lambda\xi} A\,(\phi - h(\xi))\, d\xi\, d\theta \end{aligned} \tag{2.20}$$

where we used $\int_\theta^0 \lambda e^{\lambda(\theta-\xi)}\,d\xi = 1-e^{\lambda\theta}$. Note that the right-hand side of (2.20) belongs to V^*. Thus, if we define

$$\Delta(\lambda) = \lambda^2 I + A\left(E - \int_{-r}^0 e^{\lambda\theta} g(\theta)\,d\theta\right) \in \mathcal{L}(V, V^*),$$

then $\lambda \in \rho(\mathcal{A})$ if and only if $\Delta^{-1}(\lambda) \in \mathcal{L}(V^*, V)$. For fixed $\lambda \in \mathcal{C}$ consider the sesquilinear form μ on V:

$$\begin{aligned}\mu(u,v) &= \langle \Delta(\lambda)u, v\rangle_{V^*,V} \\ &= \lambda^2\langle u, v\rangle_H + \left(E - \int_{-r}^0 e^{\lambda\theta}\, g(\theta)\,d\theta\right)\sigma(u,v) \quad \text{for } u, v \in V.\end{aligned} \tag{2.21}$$

Then, for some constant C

$$|\mu(u,v)| \le C|u|_V\,|v|_V \quad \text{for } u, v \in V,$$

and for $\lambda = a + bi$ with $a > 0$

$$\operatorname{Re}\mu(u,u) \ge (a^2 - b^2)|u|_H^2 + \alpha|u|_V^2 \quad \text{for } u \in V.$$

Thus, it follows from [15, Thm. 3.2.A] that if $\lambda > 0$, then $\Delta^{-1}(\lambda) \in \mathcal{L}(V^*, V)$.

Since Z is a Hilbert space, it follows from [13, Thm. 1.4.6] that $\operatorname{dom}(\mathcal{A})$ is dense in Z. Therefore the theorem follows from the Lumer–Phillips theorem. □

3. Approximation schemes. In this section we consider the construction of approximation schemes for (2.11)–(2.12), and we give convergence arguments based on the Trotter–Kato theorem (see [13]). Construction of these schemes involves choosing finite-dimensional subspaces of the state space Z and an approximation of the state operator $\mathcal{A}$. It can be seen that this involves two stages—the spatial variable, which includes the spaces H and V and the operator A, and the delay variable, which includes the space W and the operator D. We assume r is a fixed positive constant. For the spatial variable let V^N be any sequence of finite-dimensional subspaces of V. We assume the following approximation condition:

(C1) for any $\phi \in V$, there exists a sequence $\phi^N \in V^N$ such that $|\phi^N - \phi|_V \to 0$ as $n \to \infty$.

P_H^N and P_V^N will denote the orthogonal projection of H and V onto V^N, respectively. Let us define continuous linear operators $A^N : V^N \to V^N$ by

$$\langle A^N x, y\rangle_H = \sigma(x,y) \quad \text{for } x, y \in V^N. \tag{3.1}$$

Thus, the spaces V^N complete the discretization of the spatial variable in the sense that they give finite-dimensional subspaces of V and H, and (3.1) defines an approximation of the operator A. This is often equivalent to simply choosing a standard finite element scheme. We consider next the discretization of the delay variable. There are many approximation schemes available for delay equations, and we shall consider the so-called averaging scheme (see [1]) and a newer spline-based scheme recently developed by Ito and Kappel (see [10]). What is involved in any case is the discretization

of the delay interval $[-r, 0]$ and an approximation of the differential operator D. To proceed, let B_i^M, $i = 1, \cdots, M$ be the usual first-order spline elements corresponding to the mesh $\theta_j^M = -jr/M$, $j = 0, \cdots, M$; i.e.,

$$B_i^M(\theta) = \begin{cases} \frac{M}{r}\left(\theta - \theta_{i+1}^M\right) & \text{for } \theta_{i+1}^M \le \theta \le \theta_i^M \\ \frac{M}{r}\left(\theta_{i-1}^M - \theta\right) & \text{for } \theta_i^M \le \theta \le \theta_{i-1}^M \\ 0 & \text{elsewhere,} \end{cases}$$

and put $E_i^M = \chi_{[\theta_i^M, \theta_{i-1}^M)}$, $i = 1, \cdots, M$. Here, $\chi_{[a,b)}$ denotes the usual characteristic function on the interval $[a, b)$. For each integer N, M, we define the following subspaces of $W = L_g^2(-r, 0; V)$:

$$\widetilde{W}^{N,M} = \left\{ w \in W \mid w = \sum_{i=1}^{M} b_i^M B_i^M,\ b_i^M \in V^N \right\}$$

(3.2)

$$W^{N,M} = \left\{ w \in W \mid w = \sum_{i=1}^{M} a_i^M E_i^M,\ a_i^M \in V^N \right\}.$$

Let $P_g^{N,M}$ be the orthogonal projection of W onto $W^{N,M}$. Then we have the following convergence result for the orthogonal projections.

LEMMA 3.1. *$P_g^{N,M} h \to h$ for all $h \in W$.*

Proof. It follows from arguments used in Theorem 2.1 that D is dissipative in W and range $(I - D) = W$. Hence it follows from [13, Thm. 1.4.6] that

$$\operatorname{dom}(D) = \{w \in W : Dw \in W \text{ and } w(0) = 0\}$$

is dense in W. If $w \in \operatorname{dom}(D)$ then $g|w|_V^2$ is uniformly bounded on $[-r, 0]$. In fact, for $\theta \le \hat{\theta}$

$$\begin{aligned} g(\theta)|w(\hat{\theta})|_V^2 &= g(\theta)\Big|\int_{\hat{\theta}}^{0} Dw\, d\xi\Big|_V^2 \\ &\le \int_{\hat{\theta}}^{0} \frac{g(\theta)}{g(\xi)}\, d\xi \int_{\hat{\theta}}^{0} g|Dw|_V^2\, d\xi \le r|Dw|_W^2. \end{aligned}$$

Similarly, $\operatorname{dom}(D^2)$ is dense in W and for $\theta \le \hat{\theta}$

$$g(\theta)|Dw(\hat{\theta})|_V^2 \le r|D^2 w|_W^2 \quad \text{for } w \in \operatorname{dom}(D^2).$$

For $w \in \operatorname{dom}(D^2)$, define

$$w^{N,M}(\theta) = \sum_{i=1}^{M} w^N(\theta_{i-1}^M) E_i^M(\theta), \quad \text{where } w^N = P_V^N w.$$

Note that

$$w^N(\theta) = P_V^N w(\theta) = P_V^N \int_0^{\theta} Dw d\xi = \int_0^{\theta} P_V^N Dw d\xi.$$

Thus, $DP_V^N w = P_V^N Dw$ for $w \in \mathrm{dom}(D)$ so that if $w \in \mathrm{dom}(D^2)$, then $w^N \in \mathrm{dom}(D^2)$ and $|D^2 w^N|_W \leq |D^2 w|_W$. Hence for $w \in \mathrm{dom}(D^2)$,

$$\left|w - P_g^{N,M} w\right|_W \leq |w - w^{N,M}|_W \leq |w - w^N|_W + |w^N - w^{N,M}|_W$$

where the first term of the right-hand side tends to zero as $N \to \infty$ by the dominated convergence theorem. By the mean value theorem, for $\theta \in \left(\theta_i^M, \theta_{i-1}^M\right)$, $i = 1, \cdots, M$, there exists $\xi(\theta) \in \left(\theta, \theta_{i-1}^M\right)$ such that

$$w^N(\theta) - w^{N,M}(\theta) = Dw^N(\xi)\left(\theta - \theta_{i-1}^M\right).$$

Thus

$$\begin{aligned} |w^N - w^{N,M}|_W^2 &\leq \sum_{i=1}^M \int_{\theta_i^M}^{\theta_{i-1}^M} g(\theta)\, |Dw^N(\xi)|_V^2 \left(\theta - \theta_{i-1}^M\right)^2 d\theta \\ &\leq \frac{M}{3} r\, |D^2 w^N|_W^2 \left(\frac{r}{M}\right)^3. \end{aligned}$$

Hence for $w \in \mathrm{dom}(D^2)$

$$\left|w - P_g^{N,M} w\right|_W \to 0 \quad \text{as } N, M \to \infty.$$

Since $\mathrm{dom}(D^2)$ is dense in W and $|P_g^{N,M}| \leq 1$, this implies that $P_g^{N,M} \to I$, strongly in W. □

In order to define approximations of the operator D, consider the sesquilinear form $a^{N,M}$ on $\widetilde{W}^{N,M} \times W^{N,M}$ defined by

$$a^{N,M}\left(w^{N,M}, h^{N,M}\right) = \int_{-r}^0 \left\langle Dw^{N,M}, h^{N,M}\right\rangle_V g(\theta)\, d\theta \tag{3.3}$$

for $w^{N,M} \in \widetilde{W}^{N,M}$ and $h^{N,M} \in W^{N,M}$. Since $a^{N,M}$ is continuous, there exists a linear operator $\widetilde{D}^{N,M} : \widetilde{W}^{N,M} \to W^{N,M}$ such that

$$a^{N,M}\left(w^{N,M}, h^{N,M}\right) = \left\langle \widetilde{D}^{N,M} w^{N,M}, h^{N,M}\right\rangle_W.$$

Note that $D\widetilde{W}^{N,M} = W^{N,M}$. Hence a simple calculation shows that $\widetilde{D}^{N,M}$ is given by

$$\widetilde{D}^{N,M} w^{N,M} = \frac{M}{r} \sum_{i=1}^M \left(b_{i-1}^M - b_i^M\right) E_i^M \tag{3.4}$$

for $w^{N,M} = \sum_{i=1}^M b_i^M B_i^M$, $b_i^M \in V^N$, where $b_0^M = 0$. We will consider two isomorphisms $i_k^{N,M} : \widetilde{W}^{N,M} \to W^{N,M}$, $k = 1, 2$, defined by

$$i_1^{N,M} w^{N,M} = \sum_{i=1}^M b_i^M E_i^M \tag{3.5}$$

and

$$i_2^{N,M} w^{N,M} = \sum_{i=1}^M \frac{b_{i-1}^M + b_i^M}{2} E_i^M. \tag{3.6}$$

We define the operators $D_k^{N,M} : W^{N,M} \to W^{N,M}$ by $D_k^{N,M} = \tilde{D}^{N,M}(i_k^{N,M})^{-1}$, $k = 1, 2$. It will be apparent that the operators $D_1^{N,M}$ and $D_2^{N,M}$ correspond to the averaging scheme and the spline-based scheme, respectively. We can now complete the construction of our finite-dimensional approximation schemes for (2.11)–(2.12). Specifically, define the finite-dimensional state space $Z^{N,M} = V^N \times V^N \times W^{N,M}$, and let $P_Z^{N,M} : Z \to Z^{N,M}$ be the corresponding orthogonal projection. Letting $z^{N,M}(t) = (u^N(t), v^N(t), w^{N,M}(t))$, we consider the following differential equation on $Z^{N,M}$:

$$\frac{d}{dt} z^{N,M}(t) = \mathcal{A}_k^{N,M} z^{N,M}(t) + \operatorname{col}\left(0, P_H^N f(t), 0\right) \tag{3.7}$$

where

$$z^{N,M}(0) = \left(P_V^N u(0), P_H^N v(0), P_g^{N,M} w(0)\right).$$

Here the operator $\mathcal{A}_k^{N,M} : Z^{N,M} \to Z^{N,M}$ is defined by

$$\mathcal{A}_k^{N,M} z^{N,M} = \begin{pmatrix} v^N \\ -A^N\left(\alpha u^N + \int_{-r}^0 g(\theta) w^{N,M}(\theta)\, d\theta\right) \\ v^N + D_k^{N,M} w^{N,M} \end{pmatrix} \tag{3.8}$$

and the norm of $Z^{N,M}$ is induced from the Z-norm defined by (2.9).

In the next two subsections we prove a semigroup convergence result for each of the schemes (corresponding to $k = 1, 2$).

3.1. Averaging scheme. The isomorphism $i_1^{N,M}$ defined by (3.5) corresponds to the so-called averaging approximation considered in [1]. Recall that $S(t)$ is the C_0-semigroup generated by the operator $\mathcal{A}$. Let $S_1^{N,M}(t) : Z^{N,M} \to Z^{N,M}$ be the semigroup generated by $\mathcal{A}_1^{N,M}$. We intend to use the Trotter–Kato theorem (see [13]) to show that $S_1^{N,M}(t) P_Z^{N,M} \xrightarrow{s} S(t)$ as $N, M \to \infty$ (see Theorem 3.4 below). We will need the following two lemmas. The first is a uniform (in N and M) dissipativeness result for the operators $\mathcal{A}_1^{N,M}$, and the second is a resolvent convergence result for the operators $D_1^{N,M}$.

LEMMA 3.2. *$\mathcal{A}_1^{N,M}$ is dissipative in $Z^{N,M}$.*

Proof. For $z^{N,M} = (u^N, v^N, w^{N,M}) \in Z^{N,M}$ with $w^{N,M} = \sum_{i=1}^M a_i^M E_i^M$,

$$\begin{aligned}
\langle \mathcal{A}_1^{N,M} z^{N,M}, z^{N,M} \rangle_Z &= \alpha\sigma(u^N, v^N) - \sigma\left(\alpha u^N + \int_{-r}^0 g(\theta) w^{N,M}(\theta)\, d\theta, v^N\right) \\
&\quad + \int_{-r}^0 \left\langle v^N + \sum_{i=1}^M \frac{M}{r}\left(a_{i-1}^M - a_i^M\right) E_i^M, w^{N,M} \right\rangle_V g(\theta)\, d\theta \\
&= \sum_{i=1}^M \frac{M}{r} \left\langle a_{i-1}^M - a_i^M, a_i^M \right\rangle_V g_i^M,
\end{aligned}$$

where $a_0^M = 0$ and $g_i^M = \int_{\theta_i^M}^{\theta_{i-1}^M} g(\theta)\,d\theta$, $\quad i = 1, \cdots, M$.

Continuing, we have

$$\begin{aligned}\left\langle \mathcal{A}_1^{N,M} z^{N,M}, z^{N,M} \right\rangle_Z &= \frac{1}{2}\sum_{i=1}^{M} \frac{M}{r} g_i^M \left(\left|a_{i-1}^M\right|_V^2 - \left|a_i^M - a_{i-1}^M\right|_V^2 - \left|a_i^M\right|_V^2 \right) \\ &= -\frac{1}{2}\sum_{i=1}^{M-1} \frac{M}{r}\left(g_i^M - g_{i+1}^M\right)\left|a_i^M\right|_V^2 - \frac{1}{2}\frac{M}{r} g_M^M \left|a_M^M\right|_V^2 \\ &\quad - \frac{1}{2}\sum_{i=1}^{M} \frac{M}{r} g_i^M \left|a_i^M - a_{i-1}^M\right|_V^2 \le 0,\end{aligned}$$

since $g_i^M \ge g_{i+1}^M$ for $1 \le i \le M-1$. □

LEMMA 3.3. *For $\lambda > 0, (\lambda I - D_1^{N,M})^{-1} P_g^{N,M} h \to (\lambda I - D)^{-1} h$ for all $h \in W$.*

Proof. It follows from the arguments in Lemma 3.2 that $D_1^{N,M}$ is dissipative in $W^{N,M}$ so that $|(\lambda I - D_1^{N,M})^{-1}|_W \le 1/\lambda$ for $\lambda > 0$.

Let us define the sesquilinear form a on $\mathrm{dom}(D) \times W$ by

$$a(w,h) = \int_{-r}^{0} \langle Dw, h\rangle_V\, g(\theta)\,d\theta \quad \text{for } w \in \mathrm{dom}(D) \text{ and } h \in W. \tag{3.9}$$

For $\lambda > 0$, let $w = (\lambda I - D)^{-1} h$ and $w^{N,M} = (\lambda I - D_1^{N,M})^{-1} P_g^{N,M} h$ for $h \in W$, or equivalently $(\lambda I - D)w = h$ and $(\lambda I - D_1^{N,M}) w^{N,M} = P_g^{N,M} h$. Then, we must show that $|w - w^{N,M}|_W \to 0$ as $N, M \to \infty$. We have

$$\lambda\langle w, \gamma\rangle - a(w,\gamma) = \langle h, \gamma\rangle$$

for all $\gamma \in W$ and

$$\begin{aligned}&\lambda\left\langle i_1^{N,M} \widetilde{w}^{N,M}, \gamma^{N,M}\right\rangle - a^{N,M}\left(\widetilde{w}^{N,M}, \gamma^{N,M}\right) \\ &\qquad = \langle h, \gamma^{N,M}\rangle\end{aligned}$$

for all $\gamma^{N,M} \in W^{N,M}$, where $\widetilde{w}^{N,M} = (i_1^{N,M})^{-1} w^{N,M} = \sum_{i=1}^{M} a_i^M B_i^M$. Choosing $\gamma = \gamma^{N,M}$ in the first equation, we obtain for every $\widehat{w}^{N,M} \in \widetilde{W}^{N,M}$

$$\begin{aligned}&\lambda\left\langle i_1^{N,M}\left(\widehat{w}^{N,M} - \widetilde{w}^{N,M}\right), \gamma^{N,M}\right\rangle_W - a^{N,M}\left(\widehat{w}^{N,M} - \widetilde{w}^{N,M}, \gamma^{N,M}\right) \\ &\quad = a\left(w, \gamma^{N,M}\right) - a^{N,M}\left(\widehat{w}^{N,M}, \gamma^{N,M}\right) - \lambda\left\langle w - i_1^{N,M}\widehat{w}^{N,M}, \gamma^{N,M}\right\rangle_W\end{aligned}$$

for all $\gamma^{N,M} \in W^{N,M}$.

Let us choose $\widehat{w}^{N,M} \in \widetilde{W}^{N,M}$ as

$$\widehat{w}^{N,M} = \int_0^{\theta} P_g^{N,M}(Dw)\,d\xi. \tag{3.10}$$

Then

$$a\left(w, \gamma^{N,M}\right) - a^{N,M}\left(\widehat{w}^{N,M}, \gamma^{N,M}\right) = \int_{-r}^{0} \left\langle Dw - P_g^{N,M}(Dw), \gamma^{N,M}\right\rangle g(\theta)\,d\theta = 0$$

for all $\gamma^{N,M} \in w^{N,M}$. Using this and the fact that

$$i_1^{N,M}\left(\lambda i_1^{N,M} - \widetilde{D}^{N,M}\right)^{-1} = \left(\lambda I - D_1^{N,M}\right)^{-1} \quad \text{for } \lambda > 0,$$

it follows that

$$\left|i_1^{N,M}(\hat{w}^{N,M}-\tilde{w}^{N,M})\right|_W \le \left|P_g^{N,M}\left(w-i_1^{N,M}\hat{w}^{N,M}\right)\right|_W. \tag{3.11}$$

Moreover, we have

$$\left|w-i_1^{N,M}\hat{w}^{N,M}\right|_W \le \left|\int_0^\theta \left(Dw-P_g^{N,M}(Dw)\right)d\xi\right|_W + \left|i_1^{N,M}\hat{w}^{N,M}-\hat{w}^{N,M}\right|_W.$$

But

$$\begin{aligned}\left|\int_0^\theta \left(Dw-P_g^{N,M}(Dw)\right)d\xi\right|_W^2 &\le \int_{-r}^0 g(\theta)\int_\theta^0 \frac{d\xi}{g(\xi)}\left|Dw-P_g^{N,M}Dw\right|_W^2\\ &\le \frac{r^2}{2}\left|Dw-P_g^{N,M}Dw\right|_W^2 \to 0\end{aligned}$$

as $N, M \to \infty$. Also,

$$\begin{aligned}\left|i_1^{N,M}\hat{w}^{N,M}-\hat{w}^{N,M}\right|_W^2 &\le \sum_{i=1}^M \int_{\theta_i^M}^{\theta_{i-1}^M} g(\theta)\left(\theta-\theta_i^M\right)^2\left|\xi_i^M\right|_V^2 d\theta\\ &\le \left(\frac{r}{M}\right)^2 |Dw|_W^2 \to 0 \quad \text{as } M\to\infty\end{aligned}$$

where $P_g^{N,M}(Dw)=\sum_{i=1}^M \xi_i^M E_i^M$. Hence, $|i_1^{N,M}\hat{w}^{N,M}-w|_W \to 0$ as $N, M \to \infty$. It then follows from (3.11) that

$$\left|w^{N,M}-i_1^{N,M}\hat{w}^{N,M}\right|_W \to 0$$

so that

$$|w-w^{N,M}|_W \le \left|w^{N,M}-i_1^{N,M}\hat{w}^{N,M}\right|_W + \left|w-i_1^{N,M}\hat{w}^{N,M}\right|_W \to 0$$

as $N, M \to \infty$. □

Now we can state and prove the main result of this section.

THEOREM 3.4. *For all* $z \in Z$

$$S_1^{N,M}(t)P_Z^{N,M}z \to S(t)z \quad \text{as } N,\ M\to\infty$$

uniformly on bounded t-intervals.

Proof. In Lemma 3.2 we showed that the stability hypothesis of the Trotter–Kato theorem is satisfied. It remains for us to show convergence of $\left(\lambda I-\mathcal{A}_1^{N,M}\right)^{-1}P_Z^{N,M}z$ to $(\lambda I-\mathcal{A})^{-1}z$ for $\lambda > 0$. To do this, let $z=(\phi,\psi,h)\in Z$, and let $P_Z^{N,M}z = (\phi^N,\psi^N,h^{N,M})\in Z^{N,M}$. Also, let

$$(u,v,w)=(\lambda I-\mathcal{A})^{-1}z,$$

and

$$(u^N,v^N,w^{N,M})=\left(\lambda I-\mathcal{A}_1^{N,M}\right)^{-1}P_Z^{N,M}z.$$

Then, recalling (2.17)–(2.21), it follows that

$$w(\theta) = \int_\theta^0 e^{\lambda(\theta-\xi)}(v + h(\xi))\, d\xi, \qquad v = \lambda u - \phi,$$

and u satisfies

$$\mu(u,\chi) = \langle \psi + \lambda\phi, \chi\rangle_H + \sigma\left(\int_{-r}^0 g(\theta)(\lambda I - D)^{-1}(\phi - h(\theta))\, d\theta, \chi\right)$$

for all $\chi \in V$. Here μ is defined by (2.21). Following a similar analysis for the operator $\mathcal{A}_1^{N,M}$, we have

$$\lambda u^N - v^N = \phi^N \tag{3.12a}$$

$$\lambda v^N + A^N\left(\alpha u^N + \sum_{i=1}^M g_i^M a_i^M\right) = \psi^N \tag{3.12b}$$

$$\lambda a_i^M - v^N - \frac{M}{r}\left(a_{i-1}^M - a_i^M\right) = h_i^M, \qquad i = 1, \cdots, M \tag{3.12c}$$

where $w^{N,M} = \sum_{i=1}^M a_i^M E_i^M$ and $h^{N,M} = \sum_{i=1}^M h_i^M E_i^M$.

From (3.12c),

$$\left(\lambda + \frac{M}{r}\right) a_i^M = \frac{M}{r} a_{i-1}^M + v^N + h_i^M, \qquad i = 1, \cdots, M,$$

or

$$a_i^M = \left(1 + \frac{r}{M}\lambda\right)^{-1}\left[a_{i-1}^M + \frac{r}{M}\left(v^N + h_i^M\right)\right] \quad \text{with } a_0^M = 0.$$

By induction,

$$a_i^M = \sum_{j=1}^i \left(1 + \frac{r}{M}\lambda\right)^{j-i-1} \frac{r}{M}\left(v^N + h_j^M\right). \tag{3.13}$$

From (3.12a), $v^N = \lambda u^N - \phi^N$, and substituting these into (3.12b), we obtain

$$\lambda^2 u^N + A^N\left(\alpha u^N + \lambda \sum_{i=1}^M g_i^M \sum_{j=1}^i \left(1 + \frac{r}{M}\lambda\right)^{j-i-1} \frac{r}{M} u^N\right)$$
$$= \psi^N + \lambda\phi^N + \sum_{i=1}^M g_i^M A^N \sum_{j=1}^i \left(1 + \frac{r}{M}\lambda\right)^{j-i-1} \frac{r}{M}\left(\phi^N - h_j^M\right).$$

Here, $(1 + \frac{r}{M}\lambda)^{-1}\frac{r}{M}\lambda\sum_{j=1}^i (1 + \frac{r}{M}\lambda)^{j-i} = 1 - (1 + \frac{r}{M}\lambda)^{-i}$. Thus we have

$$\begin{aligned}\Delta_1^{N,M} u^N &= \psi^N + \lambda\phi^N \\ &\quad + A^N \int_{-r}^0 g(\theta)\left(\lambda I - D_1^{N,M}\right)^{-1}\left(\phi^N - h^{N,M}\right) d\theta\end{aligned} \tag{3.14}$$

where $\Delta_1^{N,M} \in \mathcal{L}(V^N, V^N)$ is given by

$$\Delta_1^{N,M} = \lambda^2 I + A^N\left(E - \sum_{i=1}^M g_i^M \left(1 + \frac{r}{M}\lambda\right)^{-i}\right). \tag{3.15}$$

Next note that if we define

$$e_1^M(\lambda\theta) = \sum_{i=1}^{M} \left(1 + \frac{r}{M}\lambda\right)^{-i} E_i^M(\theta), \tag{3.16}$$

then $|e^{\lambda\theta} - e_1^M(\lambda\theta)| \to 0$ as $M \to \infty$, uniformly on $[-r, 0]$ (see [11]). For each $M \geq 1$, define the sesquilinear form μ_1^M on V by

$$\mu_1^M(u, v) = \lambda^2 \langle u, v\rangle_H + \left(E - \int_{-r}^{0} g(\theta) e_1^M(\lambda\theta)\, d\theta\right) \sigma(u, v)$$

for $u, v \in V$. Then

$$\left|\mu_1^M(u, v) - \mu(u, v)\right| \leq \left(\int_{-r}^{0} g(\theta) \left|e^{\lambda\theta} - e_1^M(\lambda\theta)\right| d\theta\right) |u|_V |v|_V \tag{3.17}$$

and

$$\mu_1^M(u, u) \geq \lambda^2 |u|_H^2 + \alpha |u|_V^2 \quad \text{for all } u \in V. \tag{3.18}$$

Finally, from (3.14)–(3.15), we can write down the following equation for u^N:

$$\mu_1^M(u^N, \chi^N) = \langle \psi^N + \lambda\phi^N, \chi^N\rangle_H + \sigma\left(\int_{-r}^{0} g(\theta) \left(\lambda I - D_1^{N,M}\right)^{-1} (\phi^N - h^{N,M})\, d\theta, \chi^N\right)$$

for all $\chi^N \in V^N$.

Now, to finish the proof it is sufficient to show that $u^N \to u$. If we set $\hat{u}^N = P_V^N u$, then it is sufficient to show that $|\hat{u}^N - u^N|_V \to 0$. We have

$$\begin{aligned} \mu_1^M(\hat{u}^N - u^N, \chi^N) &= \mu_1^M(\hat{u}^N, \chi^N) - \mu_1^M(u, \chi^N) \\ &\quad + \mu_1^M(u, \chi^N) - \mu(u, \chi^N) + \lambda \langle \phi - \phi^N, \chi^N\rangle_H \\ &\quad + \sigma\left(\int_{-r}^{0} g(\theta) \left((\lambda I - D)^{-1}(\phi - h) - \left(\lambda I - D_1^{N,M}\right)^{-1} \right.\right. \\ &\qquad \left.\left. \cdot (\phi^N - h^{N,M})\right) d\theta, \chi^N\right) \end{aligned} \tag{3.19}$$

for all $\chi^N \in V^N$. It then follows from (3.17)–(3.19) that for $\lambda \geq 0$

$$\begin{aligned} \alpha |\hat{u}^N - u^N|_V &\leq \lambda^2 |u - \hat{u}^N|_H + \lambda |\phi - \phi^N|_H + K |u - \hat{u}^N|_V \\ &\quad + \left(\int_{-r}^{0} g(\theta)\, d\theta\right) \left(\sup_{\theta \in [-r,0]} \left|e^{\lambda\theta} - e_1^M(\lambda\theta)\right| |u|_V \right. \\ &\quad \left. + \left|(\lambda I - D)^{-1}(\phi - h) - \left(\lambda I - D_1^{N,M}\right)^{-1} P_g^{N,M}(\phi - h)\right|_W\right) \end{aligned}$$

where K is independent of N and M.

It thus follows from Lemma 3.3 that $|\hat{u}^N - u^N|_V \to 0$ as $N, M \to \infty$, so that for all $z \in Z$,

$$\left|\left(\lambda I - \mathcal{A}_1^{N,M}\right)^{-1} P_Z^{N,M} z - (\lambda I - \mathcal{A})^{-1} z\right|_Z \to 0$$

as $N, M \to \infty$. The result follows from the Trotter–Kato theorem. □

3.2. Spline-based scheme. The isomorphism $i_2^{N,M}$ defined by (3.6) leads to the scheme developed in [10]. Let $S_2^{N,M}(t) : Z^{N,M} \to Z^{N,M}$ be the semigroup generated by $\mathcal{A}_2^{N,M}$. The proof of convergence of $S_2^{N,M}(t)z \to S(t)z$ for $z \in Z$ will be parallel to one for $S_1^{N,M}(t)z$, given in §3.1.

LEMMA 3.5. *$\mathcal{A}_2^{N,M}$ is dissipative in $Z^{N,M}$.*

Proof. Let $z^{N,M} = (u^N, v^N, w^{N,M})$, where $w^{N,M} = \sum_{i=1}^M a_i^M E_i^M$ and where $(i_2^{N,M})^{-1} w^{N,M} = \tilde{w}^{N,M} = \sum_{i=1}^M b_i^M B_i^M$. We have

$$\begin{aligned}
&\langle \mathcal{A}_2^{N,M} z^{N,M}, z^{N,M} \rangle_Z \\
&\quad = \alpha\sigma(u^N, v^N) - \sigma\left(\alpha u^N + \int_{-r}^0 g(\theta) w^{N,M}(\theta)\, d\theta, v^N\right) \\
&\qquad + \int_{-r}^0 \left\langle v^N + \sum_{i=1}^M \left(b_{i-1}^M - b_i^M\right) \frac{M}{r} E_i^M, w^{N,M} \right\rangle_V g(\theta)\, d\theta \\
&\quad = \sum_{i=1}^M \left(\frac{M}{2r}\right) \left\langle \left(b_{i-1}^M - b_i^M\right), \left(b_{i-1}^M + b_i^M\right) \right\rangle_V g_i^M \\
&\quad = \sum_{i=1}^M \frac{1}{2} \left(\frac{M}{r}\right) g_i^M \left(\left|b_{i-1}^M\right|_V^2 - \left|b_i^M\right|_V^2 \right) \\
&\quad = -\frac{1}{2} \sum_{i=1}^{M-1} \left(\frac{M}{r}\right) \left(g_i^M - g_{i+1}^M\right) \left|b_i^M\right|_V^2 - \frac{1}{2} \left(\frac{M}{r}\right) g_M^M \left|b_M^M\right|_V^2 \le 0
\end{aligned}$$

where $a_i^M = (b_{i-1}^M + b_i^M)/2$, $1 \le i \le M$, $b_0^M = 0$, and $g_{M+1}^M = 0$. The result follows. □

Before proceeding to the main convergence result of this section, we discuss some preliminary calculations for the resolvent $\left(\lambda I - \mathcal{A}_2^{N,M}\right)^{-1}$. Consider the equation for $z^{N,M} = (u^N, v^N, w^{N,M}) \in Z^{N,M}$ given by

$$\lambda z^{N,M} - \mathcal{A}_2^{N,M} z^{N,M} = (\phi^N, \psi^N, h^{N,M}) \in Z^{N,M},$$

or equivalently

$$\lambda u^N - v^N = \phi^N \tag{3.20a}$$

$$\lambda v^N + A^N \left(\alpha u^N + \sum_{i=1}^M g_i^M a_i^M\right) = \psi^N \tag{3.20b}$$

$$\lambda \left(\frac{b_{i-1}^M + b_i^M}{2}\right) - v^N - \frac{M}{r} \left(b_{i-1}^M - b_i^M\right) = h_i^M, \quad i = 1, \cdots, M \tag{3.20c}$$

where from (3.6), $\widetilde{w}^{N,M} = \sum_{i=1}^{M} b_i^M B_i^M$ and $w^{N,M} = i_2^{N,M}\widetilde{w}^{N,M} = \sum_{i=1}^{M}((b_{i-1}^M + b_i^M)/2)\, E_i^M$ (i.e., $a_i^M = (b_{i-1}^M + b_i^M)/2, \quad i = 1, \cdots, M$) and $b_0^M = 0$. From (3.20c)

$$\left(\frac{\lambda}{2} + \frac{M}{r}\right) b_i^M = \left(\frac{M}{r} - \frac{\lambda}{2}\right) b_{i-1}^M + v^N + h_i^M,$$

or

$$b_i^M = \left(1 - \frac{r}{2M}\lambda\right)\left(1 + \frac{r}{2M}\lambda\right)^{-1} b_{i-1}^M + \frac{r}{M}\left(1 + \frac{r}{2M}\lambda\right)^{-1}(v^N + h_i^M), \qquad 1 \le i \le M.$$

By induction

$$b_i^M = \frac{r}{M}\left(1 + \frac{r}{2M}\lambda\right)^{-1} \sum_{j=1}^{i} \left(\frac{1 - \frac{r}{2M}\lambda}{1 + \frac{r}{2M}\lambda}\right)^{i-j} \left(v^N + h_j^M\right).$$

From (3.20a), $v^N = \lambda u^N - \phi^N$ and substituting these into (3.20b) we obtain

$$\lambda^2 u^N + A^N\left(Eu^N - \sum_{i=1}^{M} g_i^M \left(1 + \frac{r}{2M}\lambda\right)^{-1}\left(\frac{1 - \frac{r}{2M}\lambda}{1 + \frac{r}{2M}\lambda}\right)^{i-1} u^N\right) = \psi^N + \lambda\phi^N + A^N \int_{-r}^{0} g(\theta)\left(\lambda I - D_2^{N,M}\right)^{-1}(\phi^N - h^{N,M})\, d\theta$$

where we used

$$a_i^M = \frac{b_{i-1}^M + b_i^M}{2}$$

and

$$\frac{1}{2}\left(\sum_{j=1}^{i} x^{i-j} + \sum_{j=1}^{i-1} x^{i-1-j}\right) = \frac{1 - \left(\frac{1+x}{2}\right) x^{i-1}}{1 - x} = \left(\frac{r}{M}\lambda\right)^{-1}\left(1 + \frac{r}{2m}\lambda\right) - \left(\frac{r}{M}\lambda\right)^{-1}\left(\frac{1 - \frac{r}{2M}\lambda}{1 + \frac{r}{2M}\lambda}\right)^{i-1}, \qquad 1 \le i \le M,$$

(with $x = (1 - r\lambda/2m)/(1 + r\lambda/2m)$).

Thus, we have

$$\Delta_2^{N,M} u^N = \psi^N + \lambda\phi^N + A^N \int_{-r}^{0} g(\theta)\left(\lambda I - D_2^{N,M}\right)^{-1}(\phi^N - h^{N,M})\, d\theta \tag{3.21}$$

where $\Delta_2^{N,M} \in \mathcal{L}(V, V^*)$ is given by

$$\Delta_2^{N,M} = \lambda^2 I + A^N \left(E - \sum_{i=1}^{M} g_i^M \left(1 + \frac{r}{2M}\lambda\right)^{-1} \left(\frac{1 - \frac{r}{2M}\lambda}{1 + \frac{r}{2M}\lambda}\right)^{i-1} \right). \tag{3.22}$$

Hence, it is easy to show that $\lambda \in \rho(\mathcal{A}_2^{N,M})$ if and only if $(\Delta_2^{N,M})^{-1}$ exists. We can now state the main result of this section.

THEOREM 3.6. *For all* $z \in Z$

$$S_2^{N,M}(t) P_Z^{N,M} z \to S(t) z \quad \text{as } N, M \to \infty$$

uniformly on bounded t-intervals.

Proof. From Lemma 3.5, if $\lambda > 0$, then $\lambda \in \rho(\mathcal{A}_2^{N,M})$ and $|(\lambda I - \mathcal{A}_2^{N,M})^{-1}|_Z \leq \frac{1}{\lambda}$, $\lambda > 0$. Next, note that for $w = (\lambda I - D)^{-1} h$, $h \in W$, $\lambda > 0$,

$$\begin{aligned} \left| i_2^{N,M} \widehat{w}^{N,M} - \widehat{w}^{N,M} \right|_W^2 &\leq \sum_{i=1}^{M} \int_{\theta_i^M}^{\theta_{i-1}^M} g(\theta) \left(\theta - \frac{\theta_{i-1}^M + \theta_i^M}{2}\right) \left|\xi_i^M\right|_V^2 \, d\theta \\ &\leq \left(\frac{r}{2m}\right)^2 |Dw|_W^2 \to 0 \quad \text{as } M \to \infty \end{aligned}$$

where $\widehat{w}^{N,M}$ is given by (3.10) and $P_g^{N,M}(Dw) = \sum_{i=1}^{M} \xi_i^M E_i^M$. Hence, arguments similar to those in the proof of Lemma 3.3 allow us to show

$$\left(\lambda I - D_2^{N,M}\right)^{-1} P_g^{N,M} h \to (\lambda I - D)^{-1} h \quad \text{as } N, M \to \infty$$

for all $h \in W$. Note that for each $\lambda > 0$

$$\left| e^{\lambda\theta} - \sum_{i=1}^{M} \left(1 + \frac{r}{2m}\lambda\right)^{-1} \left(\frac{1 - \frac{r}{2m}\lambda}{1 + \frac{r}{2m}\lambda}\right)^{i-1} E_i^M \right| \to 0$$

uniformly on $[-r, 0]$. If we define the sesquilinear form μ_2^M on V by

$$\mu_2^M(u, v) = \lambda^2 \langle u, v \rangle + \left(E - \int_{-r}^{0} g(\theta) e_2^M(\lambda\theta) \, d\theta \right) \sigma(u, v)$$

for $u, v \in V$, then

$$\begin{aligned} \left| \mu_2^M(u, v) - \mu(u, v) \right| &\leq \left(\int_{-r}^{0} g(\theta) \left| e^{\lambda\theta} - e_2^M(\lambda\theta) \right| \, d\theta \right) |u|_V \, |v|_V \\ &= \epsilon_2(\lambda) |u|_V \, |v|_V \end{aligned}$$

where

$$e_2^M(\lambda\theta) = \sum_{i=1}^{M} \left(1 + \frac{r}{2m}\lambda\right)^{-1} \left(\frac{1 - \frac{r}{2m}\lambda}{1 + \frac{r}{2m}\lambda}\right)^{i-1} E_i^M$$

and $\epsilon_2(\lambda) \to 0$ as $M \to \infty$. Thus, using the exact same arguments as in the proof of Theorem 3.4, we can show that for all $z \in Z$ and $\lambda > 0$

$$\left| \left(\lambda I - \mathcal{A}_2^{N,M}\right)^{-1} P_Z^{N,M} z - (\lambda I - \mathcal{A})^{-1} z \right|_Z \to 0$$

as $N, M \to \infty$. Now, the theorem follows from the Trotter–Kato theorem. □

4. Numerical results. In this section we apply the approximation results developed above to a specific example. In particular, we consider the following model for transverse vibrations of a cantilevered linear viscoelastic beam:

$$\frac{\partial^2 u}{\partial t^2}(t,x) + \frac{\partial^2}{\partial x^2}\left[Eu_{xx}(t,x) - \int_{-r}^{0} g(s)u_{xx}(t+s,x)\,ds\right] = f(t,x) \tag{4.1}$$

where $0 < x < 1$, $t \geq 0$, and

$$u(t,0) = 0, \qquad \left[Eu_{xx}(t,x) - \int_{-r}^{0} g(s)u_{xx}(t+s,x)\,ds\right]\Bigg|_{x=1} = 0,$$

$$u_x(t,0) = 0, \qquad \frac{\partial}{\partial x}\left[Eu_{xx}(t,x) - \int_{-r}^{0} g(s)u_{xx}(t+s,x)\,ds\right]\Bigg|_{x=1} = 0.$$

Here $u(t,x)$ is the transverse displacement at time t and position x of a vibrating beam. Assume that $g(s) = \alpha e^{\beta s}/\sqrt{-s}$. It has been observed that such exponentially decaying kernels with a singularity at $s = 0$ provide a reasonable model for many viscoelastic structures (see [7], [9]). For this example we fix the parameter values in the model to be $E = 40$, $r = 1$, $\alpha = 10$, and $\beta = 5$. Define the Hilbert spaces

$$H = L_2(0,1)$$

and

$$V = \{v \in H^2(0,1) : v(0) = 0 = v'(0)\},$$

and the following sesquilinear form on $V \times V$:

$$\sigma(u,v) = \langle u'', v'' \rangle_H . \tag{4.2}$$

Then if we set $u(t) = u(t,\cdot)$, we can write (4.1) in the form of (1.1) and apply the theory we have developed for (1.1).

The first step in the approximation scheme is the construction of the finite-dimensional spaces $V^N \subset V$. For a positive integer N and partition $\{x_i\}_{i=0}^N$, $x_i = i/N$, of $[0,1]$, let S_3^N denote the set of cubic splines with knots at the x_i. Let $\{B_i^N\}_{i=1}^{N+1}$ be the standard cubic B-spline basis of S_3^N (see [14]). The set S_3^N is not contained in V, but we can define basis elements h_i^N which satisfy $h_i^N(0) = d/dx\, h_i^N(0) = 0$. That is, define

$$h_1^N = 2B_1^N - B_0^N + 2B_{-1}^N$$

and

$$h_i^N = B_i^N, \quad i = 2, \cdots, N+1.$$

Let $V^N = \operatorname{span}\{h_i^N\}_{i=1}^{N+1}$. The approximation scheme is completed with either of the delay discretizations described in §3. Hence we are led to finite-dimensional ordinary differential equations of the form

$$\dot{z}^{N,M}(t) = \mathcal{A}_k^{N,M} z^{N,M}(t) + \operatorname{col}\left(0, P_H^N f(t,x), 0\right) \tag{4.3}$$

where $k = 1$ or 2 corresponds to the choice of delay discretization.

Returning to consideration of the original equation (4.1), note that the eigenvalues of this viscoelastic system (that is, the eigenvalues of $\mathcal{A}$) are the solutions λ of the characteristic equation (recall $\Delta(\lambda)$ in §2)

$$\lambda^2 + \omega_j\left(E - \int_{-r}^{0} g(s)e^{\lambda s}\,ds\right) = 0. \tag{4.4}$$

Here $-\omega_j$, $j = 1, 2, \cdots$, are the eigenvalues of the operator A which is determined by (2.3) and (4.2). It is well known that $\omega_j = c_j^4$, where c_j, $j = 1, 2, \cdots$, are the roots of the frequency equation

$$\cos c \cdot \cosh c + 1 = 0. \tag{4.5}$$

Note that $c_j \approx (j - \frac{1}{2})\pi$ as $j \to \infty$. Note also that for the case $g(s) \equiv 0$, the solutions $\lambda_\pm$ of (4.4) are given by $\lambda_\pm = \pm i\sqrt{\omega E}$. Naturally we expect this since the case $g \equiv 0$ corresponds to no damping (i.e., perfectly elastic system). Now (4.5) and (4.4) can be solved to high accuracy numerically using Newton's method. In Table 1 we note the effect of viscoelastic damping by comparing the eigenvalues of the damped and undamped systems for the first five modes.

TABLE 1

	Undamped	Damped
Mode 1	$0 \pm 22.2372i$	$-0.69475 \pm 21.4162i$
Mode 2	$0 \pm 139.3583i$	$-1.86160 \pm 137.4791i$
Mode 3	$0 \pm 390.2074i$	$-3.13974 \pm 387.1056i$
Mode 4	$0 \pm 764.6508i$	$-4.30433 \pm 760.3019i$
Mode 5	$0 \pm 1264.0226i$	$-5.61023 \pm 1258.2092i$

Next we compare the efficacy of the two approximation schemes by observing the rate of convergence, as $N, M \to \infty$, of the eigenvalues of $A_k^{N,M}$ to the eigenvalues of $\mathcal{A}$. Actually, we fix $N = 9$, which from finite element theory gives a good approximation of the eigenvalues corresponding to the first four undamped modes. In Tables 2 and 3 we observe the behavior of these eigenvalues as M increases. We note that convergence is "frequency dependent," i.e., larger values of M are required for higher mode eigenvalues. After extensive numerical experiments, it appears that the scheme based upon the delay approximation of Ito and Kappel converges about twice as fast as the AVE scheme.

TABLE 2

	Mode 1		Mode 2	
M	AVE	SPLINE	AVE	SPLINE
4	$-.3451 + 22.189i$	$-.6052 + 22.077i$	$-0.3510 + 139.363i$	$-0.6428 + 139.344i$
16	$-.7462 + 21.913i$	$-.8493 + 21.609i$	$-0.8997 + 139.306i$	$-1.4325 + 139.168i$
64	$-.7766 + 21.527i$	$-.7339 + 21.430i$	$-1.7917 + 138.915i$	$-2.2485 + 138.291i$
400	$-.7067 + 21.430i$	$-.6977 + 21.416i$	$-2.0769 + 137.732i$	$-1.9648 + 137.526i$
1600	$-.6974 + 21.419i$	$-.6951 + 21.416i$	$-1.9177 + 137.534i$	$-1.8760 + 137.493i$
4000	$-.6957 + 21.417i$	$-.6948 + 21.416i$	$-1.8846 + 137.505i$	$-1.8654 + 137.491i$
	True $-.69475 + 21.4162i$		True $-1.86160 + 137.4791i$	

We turn finally to an idea which, from a computational point of view, may well be the most important contribution of this paper. When using the schemes discussed above for control and identification problems, large values of N and M may be required

TABLE 3

M	Mode 3 AVE	Mode 3 SPLINE	Mode 4 AVE	Mode 4 SPLINE
4	$-0.3511 + 390.517i$	$-0.6436 + 390.511i$	$-0.3512 + 767.455i$	$-0.6437 + 767.451i$
16	$-0.9042 + 390.497i$	$-1.4640 + 390.446i$	$-0.9047 + 767.444i$	$-1.4664 + 767.419i$
64	$-1.9265 + 390.339i$	$-2.8700 + 389.962i$	$-1.9431 + 767.363i$	$-2.9760 + 767.158i$
400	$-3.6268 + 388.776i$	$-3.6509 + 387.809i$	$-4.4611 + 766.203i$	$-5.2997 + 764.679i$
1600	$-3.3829 + 387.660i$	$-3.2392 + 387.415i$	$-5.0077 + 763.946i$	$-4.7064 + 763.176i$
4000	$-3.2249 + 387.491i$	$-3.1688 + 387.422i$	$-4.6596 + 763.367i$	$-4.4174 + 763.091i$
	True $-3.13974 + 387.1056i$		True $-4.30433 + 760.3019i$	

in order that the finite-dimensional system gives a good enough approximation to the infinite-dimensional system to be useful for the intended application. However, for N and M too large the computational burden may be unreasonable. Thus we were led to consider an alternative scheme in which a "nonuniform" mesh is used. More specifically, we discretize the interval $[-r, 0]$ according to the mesh θ_j^M, where $\theta_0^M = 0$, $\theta_M^M = -r$, and θ_j^M, $j = 1, 2, \cdots, M-1$ is defined so that

$$\int_{\theta_j^M}^{\theta_{j-1}^M} g(s)\, ds = \frac{c}{M}, \qquad j = 1, \cdots, M-1.$$

Here $c = \int_{-r}^{0} g(s)\, ds$. Thus the mesh is distributed according to the "mass" of the function $g(s)$. In our example this amounts to a greater concentration of mesh elements in the immediate "past history" than in the far "past history." Note that, using this mesh, we have (recall 3.15)

$$\Delta_1^{N,M} = \lambda^2 I + A^N \left(E - \sum_{i=1}^{M} \frac{c}{M} \prod_{j=1}^{i} \left(1 + \lambda \alpha_j^M\right)^{-1} \right)$$

where $\alpha_j^M = \theta_{j-1}^M - \theta_j^M$.

Because $g(s) \in L_1(-r, 0)$, it can be shown that all of the convergence arguments used for the "uniform" mesh can be suitably modified so as to hold for this nonuniform mesh. In fact, for the AVE scheme, these arguments can be found in [12]. What we gain with this new scheme, however, is a much faster rate of convergence. In Tables 4 and 5 we observe eigenvalue convergence just as in Tables 2 and 3, except now with the nonuniform mesh. Again, the scheme of Ito and Kappel performs slightly better than the AVE scheme, but the important feature is the comparison of Tables 4 and 5 with Tables 2 and 3. For example, the value $M = 8$ for the nonuniform mesh compares favorably to the value $M = 400$ for the uniform mesh. This is a significant improvement and this scheme should prove useful for the applications to control and identification mentioned above. Finally we note that the improved rate of convergence with the nonuniform mesh is as yet only a numerical observation. However, we are working on a theoretical justification of this observation.

TABLE 4

	Mode 1		Mode 2	
M	AVE	SPLINE	AVE	SPLINE
4	$-.6618 + 21.520i$	$-.7532 + 21.344i$	$-1.9313 + 138.535i$	$-2.0511 + 137.889i$
8	$-.6793 + 21.471i$	$-.7298 + 21.456i$	$-1.9081 + 137.842i$	$-1.8974 + 137.575i$
16	$-.6838 + 21.444i$	$-.6745 + 21.423i$	$-1.8683 + 137.648i$	$-1.8989 + 137.496i$
32	$-.6883 + 21.430i$	$-.6976 + 21.410i$	$-1.8633 + 137.564i$	$-1.8661 + 137.486i$
64	$-.6910 + 21.423i$	$-.6942 + 21.420i$	$-1.8636 + 137.525i$	$-1.8665 + 137.486i$
128	$-.6921 + 21.420i$	$-.6932 + 21.417i$	$-1.8645 + 137.506i$	$-1.8663 + 137.488i$
	True $-.69475 + 21.4162i$		True $-1.86160 + 137.4791i$	

TABLE 5

	Mode 3		Mode 4	
M	AVE	SPLINE	AVE	SPLINE
4	$-2.3553 + 390.123i$	$-3.4216 + 389.313i$	$-2.4200 + 767.247i$	$-3.8252 + 766.743i$
8	$-3.3751 + 388.695i$	$-3.5016 + 387.865i$	$-4.3446 + 766.046i$	$-5.1251 + 764.507i$
16	$-3.1873 + 387.879i$	$-3.1993 + 387.506i$	$-4.5922 + 764.114i$	$-4.6197 + 763.388i$
32	$-3.1358 + 387.624i$	$-3.1517 + 387.422i$	$-4.4165 + 763.545i$	$-4.4419 + 763.162i$
64	$-3.1264 + 387.510i$	$-3.1302 + 387.413i$	$-4.3830 + 763.312i$	$-4.3890 + 763.118i$
128	$-3.1250 + 387.456i$	$-3.1272 + 387.406i$	$-4.3763 + 763.205i$	$-4.3792 + 763.106i$
	True $-3.13974 + 387.1056i$		True $-4.30433 + 760.3019i$	

REFERENCES

[1] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: numerical methods based on averaging approximations*, SIAM J. Control Optim., 16 (1978), pp. 169–208.

[2] J. A. BURNS AND R. H. FABIANO, *Modeling and approximation for a viscoelastic control problem*, Distributed Parameter Systems, Lecture Notes in Control and Information Sci., 102 (1987), pp. 23–39.

[3] H. T. BANKS, R. H. FABIANO, AND Y. WANG, *Estimation of Boltzmann damping coefficients in beam models*, LCDS/CCS Report 88–13, Brown University, Providence, RI, 1988.

[4] C. M. DAFERMOS, *An abstract Volterra equation with applications to linear viscoelasticity*, J. Differential Equations, 7 (1970), pp. 554–569.

[5] ———, *Asymptotic stability in viscoelasticity*, Arch. Rational Mech. Anal., 37 (1970), pp. 297–308.

[6] ———, *Contraction semigroups and trend to equilibrium in continuum mechanics*, Symposium on Applications of Methods of Functional Analysis to Problems in Mechanics, Lecture Notes in Math., 503 (1976), pp. 295–306.

[7] W. DESCH AND R. K. MILLER, *Exponential stabilization of Volterra integrodifferential equations in Hilbert space*, J. Differential Equations, 70 (1987), pp. 366–389.

[8] W. J. HRUSA AND J. A. NOHEL, *Global existence and asymptotics in one-dimensional nonlinear viscoelasticity*, Proc. 5th Symposium on Applications of Pure Mathematics to Mechanics, Lecture Notes in Physics, 195 (1984), pp. 165–187.

[9] K. B. HANNSGEN AND R. L. WHEELER, *Time delays and boundary feedback stabilization in one-dimensional viscoelasticity*, Distributed Parameter Systems, Lecture Notes in Control and Information Sci., 102 (1987), pp. 136–152.

[10] K. ITO AND F. KAPPEL, *A uniformly differentiable approximation scheme for delay systems using splines*, Tech. Report 94-1987, Institute for Mathematics, University of Graz, Graz, Austria, 1987.

[11] I. LASIECKA AND A. MANITIUS, *Differentiability and convergence rates of approximating semigroups for retarded functional differential equations*, SIAM J. Numer. Anal., 25 (1988), pp. 883–907.

[12] R. E. MILLER, *Approximation of the LQR control problem for systems governed by partial functional differential equations*, Ph.D. dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1988.

[13] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.

[14] P. M. PRENTER, *Splines and Variational Methods*, Wiley-Interscience, New York, 1975.

[15] R. E. SHOWALTER, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, 1977.

[16] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.

[17] J. A. WALKER, *Dynamical Systems and Evolution Equations*, Plenum Press, New York, 1980.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 394-408, March 1990
©1990 Society for Industrial and Applied Mathematics
007

DIFFERENTIABILITY PROPERTIES OF SOLUTIONS OF THE EQUATION $-\varepsilon^2\Delta u + ru = f(x,y)$ IN A SQUARE*

H. HAN† AND R. B. KELLOGG‡

Abstract. A singularly-perturbed, elliptic boundary value problem is considered in a square. A theory of corner singularities for a 90° angle that takes account of lower-order terms in the equation is developed. In the case $r(x,y) \equiv$ const., an asymptotic expansion of the solution is obtained that can be differentiated termwise and which displays both the corner singularities and the boundary and corner layers of the solution.

Key words. singular perturbations, boundary layers, asymptotic [Aanalysis, corner layers

AMS(MOS) subject classifications. 35J05, 35B25

1. Introduction.

The solution of an elliptic boundary value problem possesses as much smoothness as the coefficients of the equation, the data of the problem, and the boundary of the domain permit. For example, if the domain has corners, and is therefore not smooth, the solution has corner singularities and derivatives of the solution become infinite near these corners. If the highest order terms in the equation have a small parameter that alters the balance between the ellipticity and the other terms in the equation, the solution may have a boundary layer, a region of rapid transition near the boundary of the domain. In this paper we study in a simple example a problem that contains both corner singularities and boundary layers.

We consider the equation in the title in the unit square Ω with u specified on the boundary Γ. It is known that the solution $u(x,y,\varepsilon)$ has singularities near each of the four vertices of Ω. It is also known that, for small ε, $u(x,y,\varepsilon)$ has a boundary layer behavior near Γ. We seek a joint asymptotic expansion of $u(x,y,\varepsilon)$ as $(x,y) \to$ a vertex of Ω and as $\varepsilon \to 0$ that displays the singular behavior of u in the two limits.

Butuzov [1] has already given an asymptotic expansion of u for small ε. This expansion, which serves as our starting point, contains boundary layer terms for each of the four sides of Ω, and "corner layer" terms for each of the four vertices of Ω. Butuzov has also established the uniform validity of the expansion by giving a uniform bound for the remainder in $\bar{\Omega}$. Our concern is with the differentiability of the Butuzov expansion. The importance of this lies in the fact that if the boundary data is continuous, the solution u is continuous in $\bar{\Omega}$. The corner singularities in u occur in the *derivatives* of u, starting with the second derivatives which have logarithmic singularities near the vertices of Ω. If we want to understand the interaction of the corner singularities and the boundary layer, we must obtain asymptotic expansions of the derivatives of u.

Our main result is that in the case $r(x,y) \equiv$ const.\ the Butuzov expansion can be differentiated termwise to provide an asymptotic expansion of derivatives of the

*Received by the editors September 23, 1988; accepted for publication April 10, 1989.

†Department of Applied Mathematics, Tsinghua University, Beijing, People's Republic of China, and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742. The research of this author was supported in part by the National Natural Science Foundation of China and in part by National Science Foundation grant INT-8517582.

‡Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742. The research of this author was supported in part by the Army Research Office, and in part by National Science Foundation grant INT-8517582.

solution. As a consequence of our analysis, we write the expansion in a form that shows explicitly both the traditional corner singularities and the boundary and corner layers of the solution.

In §2 we develop the Butuzov expansion. Section 3 contains a discussion of corner singularities on the square Ω. Grisvard [2] gives an exposition of this theory, especially for the Laplace operator. In this theory, there are constructed certain linear functionals of the data of a problem whose nonvanishing guarantees the presence of corner singularities in the solution of the problem. Each linear functional is associated with a vertex of the domain. In the case of the Laplace operator, when the angle of the vertex is $\pi/2$, the corresponding linear functionals are local; that is, they depend only on the data and the derivatives of the data at the vertex. We construct the appropriate linear functionals for the operator $Mu = -\Delta u + pu_x + qu_u + ru$ on the square Ω. In the construction, we pay particular attention to the effect of the lower-order terms. We find that if the coefficients p, q, r are not constant, the higher-order linear functionals are not, in general, local. In §4 we give our results on the derivatives of u.

We let $D^m u(x, y)$ denote a generic derivative of order m of u. When appropriate, we use subscripts to indicate the independent variables that are being differentiated. We let $C^{m,\alpha}(S)$ denote the usual Holder space of functions of differentiability class m, and whose mth order derivatives are Holder continuous with exponent α in a set S. When $S = \bar{\Omega}$ we simply write $C^{k,\alpha}$.

2. The Butuzov expansion. We consider the singularly perturbed elliptic boundary value problem

$$Lu \equiv -\varepsilon^2 \Delta u + r(x, y)u = f \quad \text{in } \Omega, \tag{2.1a}$$

$$u(s, 0) = g_s(s),\ u(s, 1) = g_n(s),\ u(0, s) = g_w(s),\ u(1, s) = g_e(s), \quad s \in [0, 1], \tag{2.1b}$$

where $\Omega = (0, 1) \times (0, 1)$. The functions r, f are assumed to be smooth in $\bar{\Omega}$ and $r(x, y) > 0$. We let $r_{\min} > 0$ denote the minimum value of $r(x, y)$ in $\bar{\Omega}$, and we let $a > 0$ denote a constant with $a < r_{\min}$. The boundary data g_l, h_l are assumed to be smooth in $[0, 1]$ and to satisfy the compatibility conditions $g_s(0) = g_w(0)$, $g_s(1) = g_e(0)$, etc. The asymptotic analysis of the solution of (2.1) is complicated by the fact that the domain Ω does not have a smooth boundary. Nevertheless, Butuzov [1] has given an asymptotic expansion of the solution with an arbitrary number of terms, and has provided an estimate for the remainder in the expansion. The expansion of Butuzov contains "boundary layer functions," whose role is to correct discrepancies between the boundary data and the boundary values of the reduced problem on the four sides of Ω, and "corner layer functions," whose role is to correct discrepancies between the boundary data and the boundary values of the reduced problem near the four vertices of Ω. In this section we describe the Butuzov asymptotic expansion, and we give a bound for the remainder in the expansion.

We start with the "outer expansion." For this, define $u_0(x, y) = f(x, y)/r(x, y)$, $u_1(x, y) = 0$, and $u_i(x, y) = \Delta u_{i-2}(x, y)/r(x, y)$ for $i \geq 2$. Thus, $u_i = 0$ for i odd and

$$Lu_i = -\varepsilon^2 \Delta u_i + \Delta u_{i-2}, \quad i = 2, 4, \cdots. \tag{2.2}$$

We set

$$U_{2n}(x, y) = \sum_{i=0}^{2n} \varepsilon^i u_i(x, y).$$

Then from (2.2) we have

$$LU_{2n} = f - \varepsilon^{2n+2}\Delta u_{2n}. \tag{2.3}$$

From (2.3) we see that $L(u - U_{2n}) = O(\varepsilon^{2n+2})$. Unfortunately, $u - U_{2n}$ is not small on Γ. The boundary layer functions, which we now define, are designed to correct this discrepancy in the boundary data of $u - U_{2n}$.

We introduce a stretched variable η, defined by $\eta = y/\varepsilon$. We note the formula

$$Lv(x,\eta) = -\varepsilon^2 v_{xx}(x,\eta) - v_{\eta\eta}(x,\eta) + r(x,y)v(x,\eta).$$

We make a formal expansion of the equation $Lv = 0$ into powers of ε, and equate to zero the coefficients of ε^i, to define a sequence of functions $v_0, v_1, \cdots$, as follows.

$$\begin{aligned} &-v_{0,\eta\eta} + r_0(x)v_0 = 0, \\ &-v_{1,\eta\eta} + r_0(x)v_1 = -\eta r_1(x)v_0, \\ &-v_{i,\eta\eta} + r_0(x)v_i = v_{i-2,xx} - \sum_{l=0}^{i-1} \eta^{i-l} r_{i-l}(x)v_l, \quad i = 2, \cdots, 2n. \end{aligned} \tag{2.4a}$$

The functions $r_i(x)$ are defined to be the coefficients in the Taylor series expansion of $r(x,y)$ in the variable y:

$$r(x,y) = \sum_{i=0} y^i r_i(x).$$

The equations (2.4a) are a sequence of ordinary differential equations that are used to recursively determine $v_0, v_1, \cdots$. We must specify boundary conditions to complete the determination of the v_i. We specify the boundary conditions

$$\begin{aligned} &v_0(x,0) = g_s(x) - u_0(x,0), \quad v_0(x,\eta) \to 0 \quad \text{as } \eta \to \infty, \\ &v_i(x,0) = -u_i(x,0), \quad v_i(x,\eta) \to 0 \quad \text{as } \eta \to \infty, \quad i = 1, 2, \cdots. \end{aligned} \tag{2.4b}$$

It is easy to see from the positivity of $r_0(x)$ that (2.4a,b) has a solution $v_i(x,\eta)$, and that the functions v_i satisfy

$$|D_{x\eta}^m v_i(x,\eta)| \le c_m e^{-a\eta}, \qquad m = 0, 1, \cdots, \tag{2.5}$$

where the constant c_m depends on the data of (2.1) and its derivatives. The constant $a \in (0, r_{\min})$. We also note the formulas

$$\begin{aligned} Lv_0(x,\eta) &= -\varepsilon^2 v_{0,xx}(x,\eta) + [r(x,y) - r_0(x)]v_0(x,\eta), \\ Lv_1(x,\eta) &= -\varepsilon^2 v_{1,xx}(x,\eta) + r(x,y)v_1(x,\eta) - r_0(x)v_1(x,\eta) - \eta r_1(x)v_0(x,\eta), \\ Lv_i(x,\eta) &= -\varepsilon^2 v_{i,xx}(x,\eta) + v_{i-2,xx}(x,\eta) + r(x,y)v_i(x,\eta) \\ &\quad - \sum_{l=0}^{i} \eta^{i-l} r_{i-l}(x)v_l(x,\eta), \qquad i \ge 2. \end{aligned}$$

We set

$$V_{2n}(x,\eta) = \sum_{i=0}^{2n} \varepsilon^i v_i(x,\eta).$$

We then have

$$\begin{aligned} LV_{2n}(x,\eta) = &-\varepsilon^{2n+2} v_{2n,xx}(x,\eta) - \varepsilon^{2n+1} v_{2n-1,xx}(x,\eta) \\ &+ \sum_{i=0}^{2n} \varepsilon^i v_i(x,\eta) \left\{ r(x,y) - \sum_{k=0}^{2n-i} y^k r_k(x) \right\}, \end{aligned} \tag{2.6a}$$

$$V_{2n}(x,0) = g_s(x) - U_{2n}(x,0), \tag{2.6b}$$

$$|D_{x\eta}^m V_{2n}(x,\eta)| \le c_m e^{-a\eta}, \quad m = 0,1,\cdots. \tag{2.6c}$$

In a manner similar to the construction of the v_i, we define a stretched variable $\xi = x/\varepsilon$, and we define functions $w_i(\xi,y)$ and $W_{2n}(\xi,y) = \sum_{i=0}^{2n} \varepsilon^i w_i(\xi,y)$. Also, we define stretched variables $\bar\eta = (1-y)/\varepsilon$ and $\bar\xi = (1-x)/\varepsilon$, and we define boundary layer functions $\bar v_i(x,\bar\eta)$ and $\bar w_i(\bar\xi,y)$ as above. We also define functions $\bar V_{2n}(x,\bar\eta)$ and $\bar W_{2n}(\bar\xi,y)$. These functions satisfy equations analogous to (2.6). Then we find that the remainder, $u - U_{2n} - V_{2n} - W_{2n} - \bar V_{2n} - \bar W_{2n}$, is not small near the four vertices of Ω. We introduce corner layer functions to correct this discrepancy in the boundary data near the corners.

Consider a function $z^1(\xi,\eta)$ of the stretched variables (ξ,η). Note that

$$Lz^1(\xi,\eta) = -z^1_{\xi\xi} - z^1_{\eta\eta} + r(x,y)z^1.$$

We make a formal expansion of the equation $Lz^1 = 0$ into powers of ε, and equate to zero the coefficients of ε^i, to define a sequence of functions $z_i^1(\xi,\eta)$. In defining the functions z_i^1 we shall require the Taylor series expansion of $r(\varepsilon\xi,\varepsilon\eta)$ in powers of ε. We write this expansion as

$$r(\varepsilon\xi,\varepsilon\eta) = \sum_{i=0} \varepsilon^i \rho_i^1(\xi,\eta),$$

where $\rho_0^1 = r(0,0)$, $\rho_1^1 = r_x(0,0)\xi + r_y(0,0)\eta$, and in general $\rho_i^1(\xi,\eta)$ is a homogeneous polynomial of degree i. With this, the functions $z_i^1(\xi,\eta)$ are defined by

$$\begin{aligned} &-z^1_{0,\xi\xi} - z^1_{0,\eta\eta} + \rho_0^1 z_0^1 = 0, \\ &-z^1_{i,\xi\xi} - z^1_{i,\eta\eta} + \rho_0^1 z_i^1 = -\sum_{l=0}^{i-1} \rho_{i-l}^1 z_l^1, \quad i = 1,\cdots,2n. \end{aligned} \tag{2.7a}$$

We use the boundary conditions

$$\begin{aligned} &z_i^1(\xi,0) = -w_i(\xi,0), \qquad 0 \le i \le 2n, \\ &z_i^1(0,\eta) = -v_i(0,\eta), \qquad 0 \le i \le 2n, \\ &z_i^1(\xi,\eta) \to 0 \qquad \text{as } \xi,\eta \to \infty. \end{aligned} \tag{2.7b}$$

We note that the boundary data in (2.7b) are compatible at the origin:

$$w_i(0,0) - v_i(0,0) = \begin{cases} g_s(0) - g_w(0) = 0, & i = 0, \\ 0, & i > 0. \end{cases}$$

It can be shown that (2.7a,b) has a unique solution z_i^1 which satisfies

$$|z_i^1(\xi,\eta)| \le ce^{-a\rho}, \quad 0 \le \xi,\eta \le \infty, \tag{2.8a}$$

$$|D_{\xi\eta}^m z_i^1(\xi,\eta)| \le ce^{-a\rho}, \quad \rho \ge 1, \tag{2.8b}$$

where we have set $\rho = (\xi^2+\eta^2)^{1/2}$. The restriction $\rho \ge 1$ in (2.8b) is to avoid the possible corner singularities in these functions at the origin. These corner singularities will be discussed at the end of §3. We also note the formulas

$$Lz_0^1(\xi,\eta) = [r(x,y) - \rho_0^1(\xi,\eta)]z_0^1(\xi,\eta),$$

$$Lz_i^1(\xi,\eta) = r(x,y)z_i^1(\xi,\eta) - \sum_{j=0}^{i} \rho_{i-j}^1(\xi,\eta)z_j^1(\xi,\eta).$$

We set

$$Z_{2n}^1(\xi,\eta) = \sum_{i=0}^{2n} \varepsilon^i z_i^1(\xi,\eta).$$

We then have

$$LZ_{2n}^1(\xi,\eta) = \sum_{i=0}^{2n} \varepsilon^i z_i^1(\xi,\eta) \left\{ r(x,y) - \sum_{j=0}^{2n-i} \varepsilon^j \rho_j^1(\xi,\eta) \right\}, \tag{2.9a}$$

$$Z_{2n}^1(\xi,0) = -W_{2n}(\xi,0), \qquad Z_{2n}^1(0,\eta) = -V_{2n}(0,\eta), \tag{2.9b}$$

$$|Z_{2m}^1(\xi,\eta)| \le ce^{-a(\xi+\eta)}. \tag{2.9c}$$

We also define corner layer functions $z_i^2(\bar{\xi},\eta)$, $z_i^3(\xi,\bar{\eta})$, and $z_i^4(\bar{\xi},\bar{\eta})$, respectively, at the other three corners, $(1,0)$, $(0,1)$, and $(1,1)$ of Ω. These functions satisfy equations similar to (2.7a,b), and the corresponding sums, Z_{2n}^l, $l = 2,3,4$, satisfy equations similar to (2.9).

Using the functions defined above, the asymptotic expansion $\tilde{u}_{2n}$ of Butuzov is defined to be

$$\tilde{u}_{2n} = U_{2n} + V_{2n} + W_{2n} + \bar{V}_{2n} + \bar{W}_{2n} + \sum_{l=1}^{4} Z_{2n}^l. \tag{2.10}$$

We let $R_{2n} = u - \tilde{u}_{2n}$ denote the remainder in the asymptotic expansion.

We have the theorem of Butuzov.

THEOREM 2.1. *Let u solve* (2.1a,b). *There is a constant $c_n > 0$ that is independent of ε such that*

$$|u(x,y) - \tilde{u}_{2n}(x,y)| \le c_n \varepsilon^{2n+1}.$$

Proof. We examine the boundary value problem satisfied by R_{2n}. From (2.3) we have $|L(u - U_{2n})| \le c\varepsilon^{2n+2}$. Since

$$\left| r(x,y) - \sum_{k=0}^{2n-i} y^k r_k(x) \right| \le cy^{2n-i+1} = c\varepsilon^{2n-i+1}\eta^{2n-i+1},$$

the inequalities in (2.6) give

$$|LV_{2n}(x,\eta)| \le c\varepsilon^{2n+1} + c\varepsilon^{2n+1} \sum_{i=0}^{2n} \eta^{2n-i+1} e^{-a\eta} \le c\varepsilon^{2n+1}.$$

Similar inequalities hold for the other terms that appear in the definition of $\tilde{u}_{2n}$. Hence

$$|LR_{2n}(x,y)| \le c\varepsilon^{2n+1}, \qquad (x,y) \in \Omega. \tag{2.11}$$

To examine the size of R_{2n} on $\partial\Omega$ we consider the typical side $y = 0$. From (2.6) we have

$$u(x,0) - U_{2n}(x,0) - V_{2n}(x,0) = 0.$$

From (2.9) we have

$$Z_{2n}^1(\xi,0) + W_{2n}(\xi,0) = 0,$$

and similarly,

$$Z^2_{2n}(\bar{\xi},0) + \bar{W}_{2n}(\bar{\xi},0) = 0.$$

Hence

$$R_{2n}(x,0) = -\bar{V}_{2n}(x,1/\varepsilon) - Z^3_{2n}(\xi,1/\varepsilon) - Z^4_{2n}(\bar{\xi},1/\varepsilon). \tag{2.12}$$

From the exponential decay of these functions, $|R_{2n}(x,0)| \leq c\exp(-a/\varepsilon) \leq \varepsilon^{2n+1}$. Reasoning in a similar manner on the other sides of Ω, we obtain

$$|R_{2n}(x,y)| \leq c\exp(-a/\varepsilon) \leq c\varepsilon^{2n+1}, \qquad (x,y) \in \partial\Omega. \tag{2.13}$$

Using (2.11), (2.13), and the maximum principle for the operator L, we obtain the result. □

Using the exponential decay of the boundary layer functions, we may differentiate both sides of (2.12) an arbitrary number of times with respect to x and obtain

$$|D^m_x R_{2n}(x,0)| \leq ce^{-a/\varepsilon}. \tag{2.14}$$

Similar inequalities hold on the other three sides of Ω. These inequalities will be used in §4.

It is useful to interpret the expansion (2.9) in terms of the asymptotic analysis of a function of three variables. We are concerned with the solution $u(x,y,\varepsilon)$ of (2.1) in a neighborhood of $x = y = \varepsilon = 0$. The function $\tilde{u}_{2n}$ is defined by (2.9) as the sum of functions of fewer variables: the functions u_i, which are independent of ε, the functions v_i, which depend on $(x, y/\varepsilon)$, the functions z^1_i, which depend on $(x/\varepsilon, y/\varepsilon)$, etc. The asymptotic analysis of u has thus been reduced to an asymptotic analysis of functions of fewer independent variables.

3. Compatibility conditions. On a domain with a smooth boundary, the solution of an elliptic boundary value problem has an amount of differentiability that is determined by the differentiability of the data. If the boundary of the domain is *piecewise smooth*, as for example, a polygon, this is no longer the case. In addition to the differentiability of the data, some auxiliary conditions are also required to guarantee the differentiability of the solution. These auxiliary conditions are expressed by the vanishing of certain linear functionals of the data. The theory that leads to these auxiliary conditions is presented in Grisvard [2]. In this section we present this theory in the case of the square Ω, and with special regard for the effect of the lower-order terms. We find that in the case of constant coefficients, the linear functionals are local, and the vanishing of the linear functionals may be regarded as a statement of compatibility between the boundary data at a vertex and the differential equation at that vertex. It is this fact that enables the analysis in the paper to be carried out. If the coefficients of the equation are not constant, the linear functionals are not in general local functions of the data.

We develop the linear functionals and the compatibility conditions for the boundary value problem

$$Mu \equiv -\Delta u + p(x,y)u_x + q(x,y)u_y + r(x,y)u = f \quad \text{in } \Omega, \tag{3.1a}$$

$$u(s,0) = g_s(s), u(s,1) = g_n(s), u(0,s) = g_w(s), u(1,s) = g_e(s), \quad s \in [0,1]. \tag{3.1b}$$

We suppose that the coefficients and data of the problem are smooth. We let Ω^0 be the set $\partial\Omega$ with the four vertices excluded.

A sequence of linear functionals, $\bar{\Lambda}_0(f,g), \bar{\Lambda}_1(f,g), \cdots$, is defined by

$$\bar{\Lambda}_0(f,g) = g_s(0) - g_w(0),$$
$$\bar{\Lambda}_1(f,g) = g_s''(0) + g_w''(0) + f(0,0),$$
$$\bar{\Lambda}_2(f,g) = g_s^{(4)}(0) - g_w^{(4)}(0) + D_x^2 f(0,0) - D_y^2 f(0,0),$$

and, in general,

$$\bar{\Lambda}_k(f,g) = g_s^{(2k)}(0) + (-1)^{k+1} g_w^{(2k)}(0) + \sum_{i=1}^{k} (-1)^{i-1} D_x^{2(k-i)} D_y^{2(i-1)} f(0,0). \tag{3.2}$$

We also write $\bar{\Lambda}_k = \bar{\Lambda}_k^{(1)}$, and in a similar way we define linear functionals $\bar{\Lambda}_k^{(l)}$, $l = 2,3,4$, at the other three vertices of Ω. $\bar{\Lambda}_k^{(l)}(f,g)$ is defined only if f and g possess sufficient regularity. It is convenient to define a space $X_{k,\alpha}$ of pairs (f,g) such that $f \in C^{k-2,\alpha}(\bar{\Omega})$, $g_s \in C^{k,\alpha}([0,1])$, and similarly for the other three sides of Ω. $X_{k,\alpha}$ is a Banach space with the corresponding norm. $\bar{\Lambda}_j^{(l)}$ is a bounded linear functional on $X_{2k,\alpha}$ for $j \le k$.

The linear functionals $\bar{\Lambda}_k^{(l)}$ pertain to the Poisson equation on Ω. Sometimes we shall need to apply $\bar{\Lambda}_k^{(l)}$ to the data generated by a function u. For this, we let γu denote the boundary values of u. Then $\bar{\Lambda}_k^{(l)}(-\Delta u, \gamma u)$ refers to the value of $\bar{\Lambda}_k^{(l)}$ on the data generated by the function u. Volkov [3] has analyzed the regularity of solutions of the Poisson equation, and the results are presented in Grisvard [2]. From these sources we cite the following theorem.

THEOREM 3.1. *Let u solve the problem $-\Delta u = f$ in Ω with boundary conditions* (3.1b). *Let $k \ge 1$ and suppose $(f,g) \in X_{2k,\alpha}$. Then the solution $u \in C^{2k,\alpha}(\bar{\Omega})$ if and only if*

$$\bar{\Lambda}_j^{(l)}(f,g) = 0, \qquad 0 \le j \le k, \quad 1 \le l \le 4. \tag{3.3}$$

If, furthermore, $(f,g) \in X_{2k+1,\alpha}$, the solution $u \in C^{2k+1,\alpha}(\bar{\Omega})$ if and only if (3.3) *holds.*

We want to apply this result to the problem (3.1). For this, let f and g be given, let u be the solution of (3.1), and set $F = f - ru - pu_x - qu_y$. We then define linear functionals $\Lambda_k^{(l)}(f,g)$ by

$$\Lambda_k^{(l)}(f,g) = \bar{\Lambda}_k^{(l)}(F,g).$$

As above, $\Lambda_k^{(l)}(Mu, \gamma u)$ provides the value of u on the data generated by the function u. In the case $l = 1$ we write $\Lambda_k(f,g)$ for $\Lambda_k^{(l)}(f,g)$.

The problem of determining whether $\Lambda_k(f,g)$ is well defined is a little tricky and is best settled by an inductive argument. Suppose the data $(f,g) \in X_{2,\alpha}$, and suppose further that $\bar{\Lambda}_0^{(l)}(f,g) = 0$ for $l = 1,2,3,4$. That is, suppose the boundary data are compatible at the four vertices of Ω. Then it turns out that the solution u of (3.1) lies in $C^{1,\alpha}(\bar{\Omega})$. Hence $(F,g) \in X_{2,\alpha}$, so $\Lambda_1^{(l)}(f,g) = \bar{\Lambda}_1^{(l)}(F,g)$ is a bounded linear functional on $X_{2,\alpha}$. The next linear functional, $\Lambda_2^{(l)}(f,g)$, is well defined only if $\Lambda_1^{(l)}(f,g) = 0$, and in an inductive argument, $\Lambda_k^{(l)}(f,g)$ is well defined only if $\Lambda_j^{(l)}(f,g) = 0$ for $j = 1,2,\cdots,k-1$ and $l = 1,2,3,4$. This reasoning is the content of the following theorem.

THEOREM 3.2. *Suppose* $(f,g) \in X_{2,\alpha}$, *and suppose* $\Lambda_0^{(l)}(f,g) = 0$ *for* $l = 1,2,3,4$. *Then the solution* u *of* (3.1) *lies in* $C^{1,\alpha}(\bar{\Omega})$. *If* $(f,g) \in X_{2k,\alpha}$ *and* $u \in C^{2k-1,\alpha}(\bar{\Omega})$, *the linear functionals* $\Lambda_j^{(l)}(f,g)$, $j = 1,2,\cdots,k$ *are well defined and* $\Lambda_j^{(l)}(f,g) = 0$ *for* $j = 1,2,\cdots,k-1$. *We have* $\Lambda_k^{(l)}(f,g) = 0$ *for* $l = 1,2,3,4$ *if and only if* $u \in C^{2k,\alpha}(\bar{\Omega})$. *If, further,* $(f,g) \in X_{2k+1,\alpha}$ *and* $\Lambda_k^{(l)}(f,g) = 0$ *for* $l = 1,2,3,4$, *then* $u \in C^{2k+1,\alpha}(\bar{\Omega})$.

Proof. Suppose $(f,g) \in X_{2,\alpha}$. Since the boundary data are compatible at the vertices of Ω, g may be extended in Ω to a function $\phi \in C^{2,\alpha}(\bar{\Omega})$. Then $v = u - \phi$ satisfies a problem of the form (3.1) with zero boundary data. A reflection argument shows that $v \in C^{1,\alpha}(\bar{\Omega})$, so $u \in C^{1,\alpha}(\bar{\Omega})$. Now suppose $(f,g) \in X_{2k,\alpha}$ and $u \in C^{2k-1,\alpha}(\bar{\Omega})$. Then $F \in C^{2k-2,\alpha}(\bar{\Omega})$, so $(F,g) \in X_{2k,\alpha}$, so $\Lambda_j^{(l)}(f,g)$ is well defined for $j = 1,\cdots,k$. Again since $u \in C^{2k-1,\alpha}(\bar{\Omega})$, from Theorem 3.1 we have $\Lambda_j^{(l)}(f,g) = \bar{\Lambda}_j^{(l)}(F,g) = 0$, $j = 1,\cdots,k-1$. From Theorem 3.1, $\Lambda_k^{(l)}(f,g) = \bar{\Lambda}_k^{(l)}(F,g) = 0$ for $l = 1,2,3,4$ if and only if $u \in C^{2k,\alpha}(\bar{\Omega})$. If $(f,g) \in X_{2k+1,\alpha}$ and $\Lambda_k^{(l)}(f,g) = 0$ for $l = 1,2,3,4$, then since $u \in C^{2k,\alpha}(\bar{\Omega})$, $F \in C^{2k-1,\alpha}(\bar{\Omega})$. Hence $(F,g) \in X_{2k-1,\alpha}$, so from Theorem 3.1, $u \in C^{2k+1,\alpha}(\bar{\Omega})$. □

From its definition we note that $\bar{\Lambda}_k^{(l)}(f,g)$ is a local functional of (f,g) in that the value of the functional only depends on the values of f and g and their derivatives at a certain point, the lth vertex of Ω. This is what enables the well definedness of $\bar{\Lambda}_k^{(l)}$ to depend only on the smoothness of f and g, and not on the vanishing of the preceding functionals. If M has constant coefficients the functionals $\Lambda_k^{(l)}$ can be reformulated to be local functionals, and to be well defined regardless of the vanishing of the previous functionals. This is established in Lemma 3.1.

LEMMA 3.1. *Suppose* M *has constant coefficients. Then there are bounded linear functionals* $\tilde{\Lambda}_k^{(l)}$ *on* $X_{2k,\alpha}$, $l = 1,2,3,4$, *such that* $\tilde{\Lambda}_k^{(l)}(f,g)$ *depends only on the values and derivatives of* f *and* g *at the* l*th vertex of* Ω, *and such that if* $k \geq 1$ *and* $(f,g) \in X_{2k,\alpha}$, *the solution* u *of* (3.1) *is in* $C^{2k,\alpha}(\bar{\Omega})$ *if and only if*

$$\tilde{\Lambda}_j^{(l)}(f,g) = 0, \qquad 0 \leq j \leq k, \quad 1 \leq l \leq 4. \tag{3.4}$$

If (3.4) *hold and if, furthermore,* $(f,g) \in X_{2k+1,\alpha}$, *then* $u \in C^{2k+1,\alpha}(\bar{\Omega})$.

Proof. The definition of the linear functionals $\tilde{\Lambda}_k^{(l)}$ involves a formal differentiation of (3.1). We take the case $l = 1$ and we write $\tilde{\Lambda}_k$ instead of $\tilde{\Lambda}_k^{(1)}$. Let u be the solution of (3.1) and suppose that $u \in C^{2k-1,\alpha}(\bar{\Omega})$. From Theorem 3.2, $\Lambda_j^{(l)}(f,g) = 0$ for all l and for $0 \leq j \leq k-1$. From (3.2) we see that $\Lambda_k(f,g) = \bar{\Lambda}_k(F,g)$ involves derivatives of u of order $\leq 2k-1$ of u evaluated at $(0,0)$. Also from (3.2), each derivative of even order of u occurring in $\bar{\Lambda}_k(F,g)$ is of the form $D_x^{2\mu}D_y^{2\nu}u(0,0)$. From the boundary conditions, $D_x^j u(0,0) = g_s^j(0)$, $D_y^j u(0,0) = g_w^j(0)$. By successive differentiations of (3.1a) and evaluations at $(0,0)$ we find that any odd order derivative of u, evaluated at $(0,0)$, may be written in terms of the values and derivatives of f, g_s, and g_w at $(0,0)$. Similarly, it may be shown that any even order derivative of u of the above form, when evaluated at $(0,0)$, may be written in terms of the values and derivatives of f, g_s, and g_w at $(0,0)$. If these expressions for the derivatives of u at $(0,0)$ are substituted into the formula for $\bar{\Lambda}_k(F,g)$, we obtain a new linear functional, denoted $\tilde{\Lambda}_k(f,g)$, which depends explicitly on the values and derivatives of f and g at $(0,0)$, and which is a bounded functional on $X_{2k,\alpha}$. By the construction, $\tilde{\Lambda}_k(f,g) = \Lambda_k(f,g)$ if $u \in C^{2k-1,\alpha}(\bar{\Omega})$. Hence (3.4) holds if and only if $u \in C^{2k,\alpha}(\bar{\Omega})$. □

If the coefficients of M are not constant, this reformulation of the linear functionals will not, in general, be possible. To see this, consider for example the case $l = 1$. If u satisfies (3.1a,b) and is smooth at $(0,0)$ we have $F = f - rg_s - pg'_s - qg'_w$ at (0,0), so

$$\tilde{\Lambda}_1(f,g) = g''_s(0) - p(0,0)g'_s(0) - r(0,0)g_s(0) + g''_w(0) - q(0,0)g'_w(0) + f(0,0). \tag{3.5}$$

A similar result holds for the other vertices, so $\Lambda_1^{(l)}$ can be reformulated to be a local functional, even if the coefficients are variable. To calculate $\Lambda_2^{(1)}(f,g)$ we must obtain formulas for F_{xx} and F_{yy} at $(0,0)$. In particular, we must obtain a formula for $(pu_x)_{yy} = pu_{xyy} + 2p_y u_{xy} + p_{yy}u_x$ at (0,0). There seems to be no formula for $u_{xy}(0,0)$ in terms of the data f, g_s, g_w and their derivatives evaluated at the origin. Hence the linear functional $\Lambda_2^{(1)}$ is not, in general, a local functional. If the operator M has no first derivative terms, so $p = q = 0$, then the functionals $\Lambda_2^{(l)}$ can be reformulated to be local functionals. In this case, since $F = f - ru$, if u is smooth at $(0,0)$, $F_{xx} = f_{xx} - rg''_s - 2r_x g'_s - r_{xx}g_s$ at $(0,0)$ and similarly for F_{yy}, so we may set

$$\begin{aligned}\tilde{\Lambda}_2(f,g) = {} & g_s^{(4)}(0) - g_w^{(4)}(0) + D_x^2 f(0,0) - D_y^2 f(0,0) \\ & - r(0,0)g''_s(0) - 2r_x(0,0)g'_s(0) - r_{xx}(0,0)g_s(0) \\ & - r(0,0)g''_w(0) - 2r_y(0,0)g'_w(0) - r_{yy}(0,0)g_w(0).\end{aligned} \tag{3.6}$$

In this case, however, Λ_3 involves $r_{xy}(0,0)u_{xy}(0,0)$ and so is not, in general, a local functional.

If the solution u of (3.1) does not satisfy the compatibility conditions, we can use Theorem 3.2 to obtain estimates for the derivatives of u near the corners. For this, we introduce singular functions associated with the corners. Let (r,θ) denote polar coordinates, and let

$$\psi_k(x,y) = r^{2k}\theta\cos 2k\theta + r^{2k}(\ln r)\sin 2k\theta.$$

This function arises from taking the imaginary part of the analytic function $(x + iy)^{2k}\ln(x+iy)$. Therefore we have $\psi_k(x,0) = 0, \psi_k(0,y) = (-1)^k(\pi/2)y^{2k}$ and $\Delta\psi_k = 0$. It is easily seen that $\psi_k \in C^{2k-1,\alpha}(\bar{\Omega})$ for any $\alpha < 1$ but $D^{2k}\psi_k$ is not bounded. Hence $\Lambda_i(M\psi_k, \gamma\psi_k)$ is well defined for $i \le k$, $\Lambda_i(M\psi_k,\gamma\psi_k) = 0$ for $i < k$, and $\Lambda_k(M\psi_k,\gamma\psi_k) \neq 0$. In fact, $\Lambda_k(M\psi_k,\gamma\psi_k) = \bar{\Lambda}_k(0,\gamma\psi_k) = -(\pi/2)(2k)!$. We also write $\psi_k = \psi_k^1$, and in a similar way we construct singular functions ψ_k^l, $l = 2,3,4$ at the other vertices of Ω. The only singularity of ψ_k^l is at the lth vertex of Ω. Collecting these facts and using Theorem 3.2, we obtain the following lemma.

LEMMA 3.2. *$\Lambda_i^{(j)}(M\psi_k^l, \gamma\psi_k^l)$ is well defined for $i \le k$, vanishes for $i < k$, and satisfies*

$$\Lambda_k^{(j)}(M\psi_k^l,\gamma\psi_k^l) = \begin{cases} a_k^l = -(\pi/2)(2k)!, & j = l, \\ 0, & j \neq l. \end{cases} \tag{3.7}$$

The derivatives of ψ_k^1 satisfy the inequalities

$$|D^m\psi_k^1(x,y)| \le \begin{cases} c, & m < 2k, \\ c[1 + |\ln r|], & m = 2k, \\ cr^{-(m-2k)}, & m > 2k. \end{cases}$$

and similar inequalities hold for the other vertices.

We use the singular functions to obtain bounds on the derivatives of u. This is done, in the following theorem, by establishing a recursive set of adjustments to

the solution u, subtracting off the singular part. In the case of constant coefficients, we could use the localized linear functionals $\tilde{\Lambda}_k^{(l)}$, and a recursive definition of the adjustments to the solution would not be necessary.

THEOREM 3.3. *There are bounded linear functionals $\Psi_i^{(l)}$ on $X_{2i,\alpha}$ such that if $(f,g) \in X_{2k+1,\alpha}$, if u is the solution of* (3.1), *and if v_k is defined by*

$$u = \sum_{l=1}^{4} \sum_{i=1}^{k} \Psi_i^{(l)}(f,g)\psi_i^l + v_k,$$

then $v_k \in C^{2k+1,\alpha}$ and $||v_k||_{C^{2k+1,\alpha}} \leq c||(f,g)||_{X_{2k+1,\alpha}}$. The derivatives of u of order up to $2k+1$ satisfy the inequalities

$$|D^m u(x,y)| \leq \begin{cases} c, & m = 1, \\ c[1 + |\ln r|], & m = 2, \\ cr^{-(m-2)}, & 2 < m \leq 2k+1, \end{cases} \tag{3.8}$$

which are valid in a neighborhood of the origin. Similar inequalities hold at the other vertices of Ω.

Proof. The linear functionals are defined in a recursive fashion, and the proof is by induction. From Theorem 3.2, $u \in C^{1,\alpha}(\bar{\Omega})$ and the linear functionals $\Lambda_1^{(l)}(f,g)$ are well defined. Set $v_0 = u$. Suppose, by induction, that $\Psi_j^{(l)}$ and v_j have been defined for $j = 0, \cdots, i-1$ with $v_{i-1} \in C^{2i-1,\alpha}$. Thus, $\Lambda_j^{(l)}(Mv_{i-1}, \gamma v_{i-1}) = 0$ for $j \leq i-1$ and $\Lambda_i^{(l)}(Mv_{i-1}, \gamma v_{i-1})$ is well defined. Let

$$\Psi_i^{(l)}(f,g) = \Lambda_i^{(l)}(Mv_{i-1}, \gamma v_{i-1})/a_i^l.$$

Then

$$|\Psi_i^{(l)}(f,g)| \leq c||(Mv_{i-1}, \gamma v_{i-1})||_{X_{2i,\alpha}} \leq c_1||(f,g)||_{X_{2i,\alpha}}.$$

Also,

$$v_i = v_{i-1} - \sum_{l=1}^{4} \Psi_i^{(l)}(f,g)\psi_i^l,$$

so $\Lambda_i^{(l)}(MV_i, \gamma v_i) = 0$. Since $\Lambda_j^{(l)}(M\psi_i^l, \gamma\psi_i^l) = 0$ for $j < i$, $\Lambda_j^{(l)}(Mv_i, \gamma v_i) = 0$ for $j \leq i$. Hence $v_i \in C^{2i+1,\alpha}$ and the induction is complete. Inequality (3.8) now follows from (3.7) and the bounds for $D^m v_k(x,y)$. □

4. Expansion of the derivatives. In this section, we consider the asymptotic expansion of the derivatives of u. It is found that, in the case $r(x,y) \equiv$ const., the desired asymptotic expansion of $D^m u$ can be obtained by differentiating the Butuzov expansion. The analysis proceeds by studying the derivatives of the remainder R_{2n} in the Butuzov expansion. For this, we set $M = \varepsilon^{-2}L$ and apply the compatibility conditions developed in §3. In (2.10) all the functions on the right are smooth except possibly the Z_{2n}^l, and Z_{2n}^l is smooth at all vertices of Ω except possibly the lth vertex. Hence

$$\Lambda_k^{(l)}(M\tilde{u}_{2n}, \gamma\tilde{u}_{2n}) = \Lambda_k^{(l)}(MZ_{2n}^l, \gamma Z_{2n}^l). \tag{4.1}$$

We first require some information concerning the corner layer functions z_i^l. We state the result only in the case $l = 1$.

LEMMA 4.1. *For some $c > 0$, $a > 0$,*

$$|D^m_{\xi\eta} z^1_i(\xi,\eta)| \le \begin{cases} ce^{-a\rho}, & m = 1, \\ c[1 + |\ln \rho|]e^{-a\rho}, & m = 2, \\ c\rho^{-(m-2)}e^{-a\rho} & m \ge 3. \end{cases}$$

If $r(x,y) \equiv$ const., there are constants a_{ik} such that $\tilde{\Lambda}_k(Mz^1_i, \gamma z^1_i) = a_{ik}/\varepsilon^{2k}$.

Proof. From (2.7) we know that z^1_i is smooth for $\xi > 0$, $\eta > 0$, and the only possible singularity is at the origin. We consider z^1_i as being defined by a boundary value problem in the unit square. Using (3.8) to estimate $D^m z^1_i$ for $\rho \le 1$, and using (2.8b) for $\rho > 1$, we obtain the first assertion. Now suppose $r(x,y) \equiv$ const. Let $f_i(\xi,\eta)$ denote the right side of (2.7a), and let $F_i = Mz_i$. Then $F_i = \varepsilon^{-2}Lz_i = -z_{i,xx} - z_{i,yy} + \varepsilon^{-2} r z_i = -\varepsilon^{-2} f_i$. Note also that $D^k_x z_i(x,0) \sim \varepsilon^{-k}$, $D^k_y z_i(0,y)) \sim \varepsilon^{-k}$. A computation then shows that $\tilde{\Lambda}_k(Mz^1_i, \gamma z^1_i)$ is a homogeneous function of ε, and the degree of homogeneity is ε^{-2k}, so we obtain the second assertion. □

Although in general the derivatives of R_{2n} are not bounded in $\bar{\Omega}$, the derivatives of LR_{2n} are bounded in $\bar{\Omega}$. This fact is important in our analysis. The next lemma provides bounds for these derivatives.

LEMMA 4.2. *Suppose $r(x,y) \equiv$ const. For $m \le 2n+1$, $|D^m_{xy} LR_{2n}| \le c\varepsilon^{2n+1-m}$.*

Proof. From (2.11) the result is true for $m = 0$. From (2.3), $|D^m_{xy} L(u - U_{2n})| \le c\varepsilon^{2n+2}$. Since $D_y = \varepsilon^{-1} D_\eta$, $|D^m_{xy} v(x,\eta)| \le c\varepsilon^{-m} |D^m_{x\eta} v(x,\eta)|$. Using (2.6), we get

$$|D^m_{xy} LV_{2n}(x,\eta)| \le c\varepsilon^{2n+1-m}, \qquad m \le 2n+1.$$

Similar results hold for the other boundary layers. From (2.9), $LZ_{2n} = 0$, and similarly for the other corner layers. Combining these estimates, we obtain the result. □

We now have the following lemma.

LEMMA 4.3. *Suppose $r \equiv$ const. Then if $2k < 2n+1$, $\Lambda^{(l)}_k(MR_{2n}, \gamma R_{2n}) = 0$.*

Proof. We give the proof in the case $l = 1$. Since $u = \tilde{u}_{2n} + R_{2n}$, we use (4.1) to get

$$\varepsilon^{k+1}\Lambda_k(Mu, \gamma u) = \varepsilon^{k+1}\Lambda_k(MZ^1_{2n}, \gamma Z^1_{2n}) + \varepsilon^{k+1}\Lambda_k(MR_{2n}, \gamma R_{2n}). \tag{4.3}$$

Since $r \equiv$ const., we may use the linear functional $\tilde{\Lambda}_k$ given by Lemma 3.1. A computation shows that $\tilde{\Lambda}_k(Mu, \gamma u)$ is a polynomial in ε^{-1} of degree $\le k+1$, and from Lemma 3.1 we note that the coefficients of the polynomial are linear combinations of the derivatives $D^i_x g_s(0)$ and $D^i_y g_w(0)$ for $0 \le i \le 2k$, and the derivatives $D^i_{xy} f(0,0)$ for $0 \le i \le 2(k-1)$. Furthermore, each derivative $D^i_{xy} f(0,0)$ that appears in $\tilde{\Lambda}_k(f,g)$ appears in the coefficient of $\varepsilon^{-(2k-i)}$. Hence the left side of (4.3) is a polynomial of degree $\le k+1$ in ε with coefficients that depend only on the data f and g. From Lemma 4.1, $\varepsilon^{k+1}\tilde{\Lambda}_k(MZ^1_{2n}, \gamma Z^1_{2n})$ is a sum of powers ε^j with $-k+1 \le j \le 2n-k+1$. The coefficients of ε^j in this expression depend only on the data f and g. We now estimate $h_k(\varepsilon) = \varepsilon^{k+1}\tilde{\Lambda}_k(MR_{2n}, \gamma R_{2n})$. As we have seen, h_k may be written as a polynomial in ε with coefficients depending on ε. These coefficients are linear combinations of the derivatives $D^i_x R_{2n}(x,0)$ evaluated at $x = 0$, $0 \le i \le 2k$, the derivatives $D^i_y R_{2n}(0,y)$ evaluated at $y = 0$, $0 \le i \le 2k$, and the derivatives $D^i_{xy} LR_{2n}(0,0)$, $0 \le i \le 2(k-1)$. From (2.14) and the analogous inequalities on the other sides of Ω, the derivatives of the boundary values of R_{2n} are exponentially small. The derivatives of $LR_{2n}(0,0)$ that appear in $h_k(\varepsilon)$ appear in the form $\varepsilon^{i-k+1} D^i_{xy} LR_{2n}$, and from Lemma 4.2, these terms are bounded by $c\varepsilon^{2n+2-k}$. Hence we have proved

$$|h_k(\varepsilon)| \le c\varepsilon^{2n+2-k}. \tag{4.4}$$

Hence if $k+1 < 2n+2-k$, $h_k(\varepsilon)$ vanishes more quickly than the other two terms appearing in (4.3). Since h_k is an analytic function of ε, we must have $h_k(\varepsilon) \equiv 0$. The proof in the case of other l is similar. □

We now have the following lemma.

LEMMA 4.4. *Let κ, μ, and ν be nonnegative integers with $\mu+\nu \le m$ and $2\kappa \le \mu$. Then $\varepsilon^2 D_x^\mu D_y^\nu R$ may be written as a linear combination of $\varepsilon^2 D_x^{\mu-2\kappa} D_y^{\nu+2\kappa} R$, derivatives of R of order $\le \mu+\nu-2$, and derivatives of LR of order $\le \mu+\nu-2$. The coefficients appearing in the linear combination depend only on r and its derivatives, and in particular are independent of ε.*

Proof. The proof follows from successive differentiations of the equation $-\varepsilon^2 R_{xx} - \varepsilon^2 R_{yy} + rR = LR$ and an examination of the linear combinations of derivatives of R of order $\mu+\nu$ that are obtained. □

The next lemma gives bounds for the solution of a mixed boundary value problem associated with the operator L, with the dependence of the constants on ε made explicit.

LEMMA 4.5. *Let $R \in C^2(\bar{\Omega})$, and let $c > 0$ be such that $|LR(x,y)| \le c$ on $\bar{\Omega}$, and on each of the four sides of Ω either $R \le c$ or $|\partial R/\partial n| \le c/\varepsilon$. Then $|R(x,y)| \le 8c \max\{1, 1/r_{\min}\}$, where $r_{\min} > 0$ denotes the minimum value of r.*

Proof. Set

$$T(x,y) = 2c[4 - e^{-x/\varepsilon} - e^{-(1-x)/\varepsilon} - e^{-y/\varepsilon} - e^{-(1-y)/\varepsilon}].$$

Then a computation shows that $LT \ge cr_{\min}$ in $\bar{\Omega}$, $T(x,0) \ge c$, $T_y(x,0) \ge c/\varepsilon$, and similarly for the other three sides of Ω. Applying the maximum principle to the function $\max\{1, 1/r_{\min}\}T \pm R$, we obtain the result. □

Using these lemmas, we give the desired results for the derivatives of $\tilde{u}_{2n}$.

THEOREM 4.1. *Suppose $r(x,y) \equiv$ const., and let u be the solution of (2.1). Let $n \ge 2$. Then there is a constant $c > 0$, independent of ε, such that for $m \le 2n+1$,*

$$|D_{xy}^m R_{2n}(x,y)| = |D_{xy}^m u(x,y) - D_{xy}^m \tilde{u}_{2n}(x,y)| \le c\varepsilon^{2n+1-m}.$$

Proof. From Lemma 4.3, $\Lambda_k^{(l)}(MR_{2n}, \gamma R_{2n}) = 0$ for $k \le n$. Hence from Theorem 3.2, $R_{2n} \in C^{2n+1,\alpha}(\bar{\Omega})$, so $D_{xy}^m R_{2n}$ is continuous in $\bar{\Omega}$ for $m \le 2n+1$. The proof is by induction on m. From Theorem 2.1 the result is true for $m = 0$. To continue, let $m > 0$ and suppose the inequality is true for all orders of differentiation less than m. Let $S = D_{xy}^m R_{2n}$. To estimate S we will apply the maximum principle to a boundary value problem satisfied by S. For this we must estimate LS and the right-hand sides in the boundary conditions that we will impose on S. It is easily seen that $D_{xy}^m LR_{2n} - LD_{xy}^m R_{2n}$ is a linear combination of derivatives of R_{2n} of order less than m. Hence, by induction and Lemma 4.2,

$$|LS(x,y)| \le c\varepsilon^{2n+1-m}. \tag{4.5}$$

The boundary conditions which we impose on S depend on the particular derivative that is being estimated. Note first that from (2.14) and the analogous inequalities on the other three sides of Ω,

$$\begin{aligned} &|D_x^k R_{2n}(x,0)| \le ce^{-a/\varepsilon}, \qquad |D_x^k R_{2n}(x,1)| \le ce^{-a/\varepsilon}, \\ &|D_y^k R_{2n}(0,y)| \le ce^{-a/\varepsilon}, \qquad |D_y^k R_{2n}(1,y)| \le ce^{-a/\varepsilon}. \end{aligned} \tag{4.6}$$

Let $S = D_x^\mu D_y^\nu R_{2n}$, where $\mu+\nu = m$. The boundary value problem that we associate with S depends on the parity of μ and ν. Suppose, for example, that μ is even and

ν is odd, so m is odd. In this case, for the boundary value problem satisfied by S we impose Dirichlet conditions on the vertical sides of Ω and Neumann conditions on the horizontal sides of Ω. Using Lemma 4.4, we write S as a linear combination of $D_y^m R_{2n}$, and of derivatives of R_{2n} and LR_{2n} of order $\leq m-2$, where the coefficients of the latter derivatives are bounded by $c\varepsilon^{-2}$. We evaluate this relation on a vertical side of Ω. We use the inequality (4.6) to bound $D_y^m R_{2n}$, and (4.5) and the inductive assumption to bound the other terms in the linear combination. We obtain

$$|S(0,y)| \leq c\varepsilon^{2n+1-m}, \qquad |S(1,y)| \leq c\varepsilon^{2n+1-m}. \tag{4.7}$$

Similarly, we write $S_y = D_x^\mu D_y^{\nu+1} R_{2n}$ as a linear combination of $D_x^{m+1} R_{2n}$, and of derivatives of R_{2n} and LR_{2n} of order $\leq m-1$, where the coefficients of the latter derivatives are bounded by $c\varepsilon^{-2}$. We evaluate this relation on a horizontal side of Ω. We use the inequality (4.6) to bound $D_x^{m+1} R_{2n}$, and (4.5) and the inductive assumption to bound the other terms in the linear combination. We obtain

$$|S_y(x,0)| \leq c\varepsilon^{2n-m}, \qquad |S_y(x,1)| \leq c\varepsilon^{2n-m}. \tag{4.8}$$

Using (4.5), (4.7), (4.8), and Lemma 4.5, we obtain $|S(x,y)| \leq c\varepsilon^{2n+1-m}$. If μ is odd and ν is even, we assign Dirichlet conditions to the horizontal sides of Ω and Neumann conditions to the vertical sides of Ω. If μ and ν are both even, Dirichlet conditions are assigned to all the sides of Ω, and if μ and ν are both odd, Neumann conditions are assigned to all the sides of Ω. The reasoning is similar in all these cases. If $r(x,y) \not\equiv$ const., we use the reformulation of Λ_k for $k = 1, 2$ given in (3.5) and (3.6). From these formulas we obtain Lemma 4.1 for $k = 1, 2$, and then we obtain Lemma 4.3. The rest of the proof of (4.5) follows as before, so in the general case we obtain (4.5) for $k = 1, 2$. □

It is natural to ask whether Theorem 4.1 holds in the case $r(x,y) \not\equiv$ const. We conjecture that the result holds for $m \leq 5$, but that for $m > 5$ the Butuzov expansion must be modified to obtain an asymptotic expansion of $D^m u$. The question is not studied further here.

It is of interest to combine the corner expansion (3.7) given in §3 with the boundary layer expansion provided by (2.10). Since the outer expansion U_{2n} is smooth, and the boundary layer expansions V_{2n}, W_{2n}, etc., are smooth, and since a consequence of Lemma 4.3 is that R_{2n} is smooth, it follows that the corner singularities of u at the lth vertex of Ω are contained in the corner layer terms z_i^l of $\tilde{u}_{2n}$. To analyze these corner singularities we may use the expansion (3.7). It is useful here to introduce singular functions that are specifically adapted to the operator N that defines the corner layer functions:

$$Nz(\xi,\eta) \equiv -z_{\xi\xi}(\xi,\eta) - z_{\eta\eta}(\xi,\eta) + r(0,0)z(\xi,\eta).$$

We define singular functions $\chi_k(\xi,\eta)$ as the solution to the problem

$$\begin{aligned} &-\Delta\chi_k + r(0,0)\chi_k = 0, \qquad \xi > 0, \quad \eta > 0, \\ &\chi_k(\xi,0) = 0, \qquad \chi_k(0,\eta) = \eta^{2k}, \\ &\chi_k \to 0 \quad \text{as } \rho \to \infty. \end{aligned} \tag{4.9}$$

Considered as a function on the unit square in the $\xi\eta$ plane, χ_k has a singularity only at the origin and

$$\Lambda_i(N\chi_k, \gamma\chi) \text{ is } \begin{cases} = 0, & i < k, \\ \neq 0, & i = k, \\ \text{not defined}, & i > k. \end{cases} \tag{4.10}$$

It may be shown that χ_k satisfies the inequalities

$$|D^l\chi_k(\xi,\eta)| \le \begin{cases} ce^{-r(0,0)\rho}, & l < 2k, \\ c[1+|\ln\rho|]e^{-r(0,0)\rho}, & l = 2k, \\ c\rho^{-(l-2k)}e^{-r(0,0)\rho}, & l > 2k. \end{cases} \tag{4.11}$$

Using these functions, we write the singular expansion of $z_i^1(\xi,\eta)$ as

$$z_i^1(\xi,\eta) = \sum_{k=1}^{m} \Lambda_k(Nz_i^1, \gamma z_i^1)\chi_k(\xi,\eta) + \zeta_m^1(\xi,\eta). \tag{4.12}$$

The remainder, ζ_m^1 has $2m+1$ continuous derivatives in the positive quadrant $\xi \ge 0$, $\eta \ge 0$ and satisfies

$$|D^l\zeta_k^1(\xi,\eta)| \le ce^{-r(0,0)\rho}, \qquad 0 \le l \le 2m+1. \tag{4.13}$$

We also write $\chi_k^{(1)}(\xi,\eta) = \chi_k(\xi,\eta)$, and in a similar manner we define singular functions $\chi_k^{(l)}$ pertaining to the other vertices of Ω.

We now formulate a result that expresses the interaction between the corner singularities of the problem and the singular perturbation behavior of the problem.

THEOREM 4.2. *Let u be the solution of* (2.1). *Suppose $r(x,y) \equiv$ const. Then*

$$u(x,y) = \sum_{l=1}^{4}\sum_{k=1}^{m} a_k^l(\varepsilon)\chi_k^{(l)} + V_{2n} + W_{2n} + \bar{V}_{2n} + \bar{W}_{2n} + S_m(x,y,\varepsilon), \tag{4.14}$$

where $a_k(\varepsilon)$ is a polynomial in ε of degree $2n$, and where the remainder S_m satisfies

$$|D^l S_m(x,y,\varepsilon)| \le c, \qquad 0 \le l \le 2m+1.$$

Proof. Inserting (4.12) into (2.10) we obtain (4.14) where

$$a_k^l(\varepsilon) = \sum_{i=0}^{2n} \varepsilon^i \Lambda_k^{(l)}(Nz_i^l, \gamma z_i^l),$$

$$S_j(x,y,\varepsilon) = U_{2n} + R_{2n} + \sum_{i,l} \varepsilon^i \zeta_i^l.$$

The inequality then follows from the properties of the functions that form S_m. □

As a corollary of Theorem 4.2, we obtain bounds on the derivatives of u that display both the parameter ε and the distance to the nearest corner of Ω. To state these bounds we define distance functions d_v and d_s measuring the distance to the nearest vertex of Ω and the nearest side of $\bar{\Omega}$ respectively by

$$d_v(x,y)^2 = \min\{x^2+y^2,\ (1-x)^2+y^2,\ x^2+(1-y)^2,\ (1-x)^2+(1-y)^2\}$$

$$d_s(x,y) = \min\{x, y, 1-x, 1-y\}.$$

We then have the following corollary.

COROLLARY. *Let u be the solution of* (2.1). *Suppose that $r(x,y) \equiv$ const. Then we have*

$$|D_{xy}^m u(x,y)| \le c + c\varepsilon^{-m}e^{-ad_s(x,y)/\varepsilon} + \begin{cases} ce^{-ad_v/\varepsilon}, & m = 1, \\ c[1+|\ln(d_v/\varepsilon)|]e^{-ad_v/\varepsilon}, & m = 2, \\ c(d_v/\varepsilon)^{-(m-2)}e^{-ad_v/\varepsilon} & m \ge 3. \end{cases}$$

Proof. The result follows from Theorem 4.2 and the bounds on the derivatives of the various functions that have been established. □

REFERENCES

[1] V. F. BUTUZOV, *The asymptotic properties of solutions of the equation* $\mu^2 \Delta u - k^2(x,y)u = f(x,y)$ *in a rectangle*, Differentsial'nye Uravneniya, 9 (1973), pp. 1274–1279.

[2] P. GRISVARD, *Elliptic problems in nonsmooth domains*, Pitman, Boston, 1985.

[3] E. A. VOLKOV, *On the differential properties of solutions of boundary value problems for the Laplace and Poisson equations on a rectangle*, Trudy Mat. Inst. Steklov, 77 (1965), pp. 89–112.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 409-435, March 1990
© 1990 Society for Industrial and Applied Mathematics
008

CONVEX FUNCTION OF A MEASURE OBTAINED BY HOMOGENIZATION*

FRANÇOISE DEMENGEL† AND TANG QI‡

Abstract. This paper defines and studies convex functions of a homogenized measure. These types of mathematical objects are Γ limit of integral functionals which depend on the deformation of the displacement of a material. This work has been done in view of its applications to homogenization in perfect plasticity.

Key words. integration theory via linear functionals, operations with distributions, yield criteria

AMS(MOS) subject classifications. 28C05, 46F10, 73E05

Introduction. This paper states mathematical results necessary for studying the homogenization for one kind of material, namely, elastic perfect plastic, for Hencky's law.

The theory of homogenization consists in assuming that the domain Ω occupied by the considered material is composed of a great number of identical cells of size ε. For $\varepsilon > 0$, P_ε is defined as the prehomogenized corresponding displacement problem, and P_ε^* as the stress problem. By letting ε tend to zero the domain Ω becomes almost homogeneous and the question is, in what sense P_ε converges to the homogenized problem, which will be defined later.

Variational mechanical problems for nonhomogeneous materials other than plasticity have been studied with theories of homogenization (cf. [3], [4]), but the functionals used were coercives on reflexive spaces. The main difficulty here lies in the fact that the "prehomogenized" energies are bounded only in L^1. This has motivated this first theoretical part. The homogenized energy $\int \bar{\Psi}(u)$ is defined not only as a functional, but also as the integral on the open set Ω *of a measure.* We show in § 1 that $\bar{\Psi}(u)$ is a measure that is absolutely continuous with respect to the measure $|e^D(u)|+|\operatorname{div} u|^2$. As Buttazo and Dal Mazo [4] did for homogenization of coercive problems on H^1, we state in § 2, the existence of a punctual function, called homogenized and denoted Ψ_{hom}, such that

$$\langle \bar{\Psi}(u), \varphi \rangle = \int \Psi_{\text{hom}}(e(u)(x))\varphi(x)\, dx,$$

when u is sufficiently regular (for example, for u in $H^1(\Omega)$). Moreover, Ψ_{hom} satisfies

$$c_0(|\xi^D|+|\operatorname{tr}\xi|^2-1) \leqq \Psi_{\text{hom}}(\xi) \leqq c_1(|\xi^D|+|\operatorname{tr}\xi|^2+1).$$

The theory of convex function of a measure, studied in [11] and [12], then allows the measure $\Psi_{\text{hom}}(e(u))$ to be defined when u belongs to $U(\Omega)=\{u \in L^1(\Omega),\ (u_{i,j}+u_{j,i})/2 \in M^1(\Omega)$, for all $i,j \in [1, N]$ and div $u \in L^2(\Omega)\}$. The central result of this paper then consists of showing that the measure $\Psi_{\text{hom}}(e(u))$ *coincides with* $\bar{\Psi}(u)$; in other words, the formula just above may be extended in a certain sense to functions of $U(\Omega)$:

$$\langle \bar{\Psi}(u), \varphi \rangle = \langle \Psi_{\text{hom}}(e(u)), \varphi \rangle.$$

This result will be crucial for stating (see [9]) the convergence of prehomogenized problems towards the homogenized one.

* Received by the editors February 5, 1987; accepted for publication (in revised form) April 17, 1989.
†‡ Université Paris-Sud et Centre National de la Recherche Scientifique, Laboratoire d'Analyse Numérique d'Orsay, Bâtiment 425, 91405 Orsay Cedex, France.

We conclude by giving an "explicit" expression of Ψ_{hom}, and by determining the convex set K_{hom}, which is the domain of $(\Psi_{\text{hom}})^* = (\Psi^*)_{\text{hom}}$. These results have been obtained before by Bouchitté [3], [2] for the gradient and by Tang Qi [19] for the strains, but our method seems to be more direct and uses the properties of convex functions of a measure.

1. The homogenized measure $\bar{\Psi}(u)$ for u in $U(\Omega)$. Let Ω be an open bounded set of $\mathbb{R}^N$, $N \geqq 2$, and let E be the space of symmetric tensors of order two on $\mathbb{R}^N$. When ξ belongs to E, its deviator ξ^D is defined by $\xi^D = \xi - (\text{tr } \xi / N)\,\text{Id}$, where $\text{tr } \xi = \sum_{i=1}^N \xi_{ii}$. We denote by Y the open N-simplex $]0, 1[^N$, and we suppose that Ψ is measurable and Y-periodic with respect to the first variable, convex in relation to the second one, and verifying that

$$\Psi \text{ is nonnegative}, \qquad \Psi(x, 0) = 0. \tag{1.1}$$

There exist some positive constants c_0, c_1, such that for every $\xi \in E$, for all $x \in \mathbb{R}^N$,

$$c_0(|\xi^D| + (\text{tr } \xi)^2 - 1) \leqq \Psi(x, \xi) \leqq c_1(|\xi^D| + (\text{tr } \xi)^2 + 1). \tag{1.2}$$

A useful consequence of (1.2) and the convexity of Ψ is the following. If $\text{tr } \xi = \text{tr } \eta$, then

$$|\Psi(x, \xi) - \Psi(x, \eta)| \leqq C_1(|\xi^d - \eta^d|). \tag{1.3}$$

We denote for a real $\varepsilon > 0$ by Ψ_ε the function

$$\Psi_\varepsilon(x, \varepsilon) = \Psi\left(\frac{x}{\varepsilon}, \xi\right). \tag{1.4}$$

Now we give more notation. $e(u)$ will denote for u in L^1 the deformation of u, i.e.,

$$e(u)_{ij} = \frac{u_{i,j} + u_{j,i}}{2} \quad \forall i, j \in [1, N] \tag{1.5}$$

(where the derivatives are taken in the sense of distributions). We define the spaces:

$$LU(\Omega) = \{u \in L^1(\Omega),\ e(u) \in L^1(\Omega, E),\ \text{div } u \in L^2(\Omega)\}, \tag{1.6}$$

$$U(\Omega) = \{u \in L^1(\Omega),\ e(u) \in M^1(\Omega, E),\ \text{div } u \in L^2(\Omega)\} \tag{1.7}$$

where $M^1(\Omega, E)$ denotes the space of bounded measures on Ω that take their values in E. Let us finally recall the characterization of Γ-limits (cf. [7], [8], [1]). If (X, τ) is a topological space satisfying the first axiom of countability, F_h is a sequence of functions from X into $\bar{\mathbb{R}}$, and u is an element of X, then

$$\lambda^- = \Gamma^- \underline{\lim}_{\substack{h \to 0 \\ v \to u}} F_h(v)$$

if and only if, for every sequence u_h converging to u in (X, τ),

(i) $\lambda^- \leqq \underline{\lim}\, F_h(u_h)$;

(ii) There exists a sequence u_h converging to u in (X, τ) such that $\lambda^- = \underline{\lim}\, F_h(u_h)$.

In a similar manner we define the $\Gamma\,\overline{\lim}$, by replacing $\underline{\lim}$ by $\overline{\lim}$ in the inequalities above. It is easy to see that $\Gamma^-\,\underline{\lim}\, F_h$ and $\Gamma^-\,\overline{\lim}\, F_h$ exist, since they may also be defined as

$$\Gamma^-\,\underline{\lim}\, F_h(u) = \inf_{v_h \to u} \left\{ \underline{\lim}_{h \to 0} F_h(v_h) \right\},$$

$$\Gamma^-\,\overline{\lim}\, F_h(u) = \inf_{v_h \to u} \left\{ \overline{\lim}_{h \to 0} F_h(v_h) \right\}.$$

A sequence F_h is said to be Γ convergent in u if and only if $(\Gamma^- \underline{\lim})\, F_h(u) = (\Gamma^- \overline{\lim}\, F_h)(u)$.

In the following definitions we will take $X = L^p(\Omega)$, where p is fixed in $]1, N/N-1[$ endowed with its strong topology.

Let us take φ in $\mathscr{C}(\bar{\Omega})$,[1] $\varphi \geqq 0$, u in $U(\Omega)$, and define the functional

$$F_\varepsilon(v, \varphi) = \begin{cases} \displaystyle\int \Psi\left(\frac{x}{\varepsilon}, e(v)\right)\varphi(x)\, dx & \text{if } v \in LU(\Omega), \\ +\infty & \text{if not} \end{cases} \tag{1.8}$$

[2]

and the following Γ-limit:

$$\langle \bar{\Psi}^-(u), \varphi\rangle = \Gamma^-(L^p) \underline{\lim_{\substack{\varepsilon \mapsto 0 \\ v \mapsto u}}} F_\varepsilon(v, \varphi) < +\infty, \qquad \langle \bar{\Psi}^+(u), \varphi\rangle = \Gamma^- \overline{\lim}\, F_\varepsilon(v, \varphi). \tag{1.9}$$

In the result above (Theorem 1.1(i)) we will see that the Γ limit of $F_\varepsilon(u, \varphi)$ exists for a subsequence, for every u, and $\varphi \geqq 0$, in $C(\bar{\Omega})$. We will then restrict to this subsequence and denote by $\langle \bar{\Psi}(u), \varphi\rangle$ the corresponding Γ limit.

THEOREM 1.1. (i) *There exists a subsequence ε' for which $\bar{\Psi}^-(u) = \bar{\Psi}^+(u) = \bar{\Psi}(u)$ for every u in $U(\Omega)$. Moreover, $\bar{\Psi}(u)$ may be extended as a bounded measure on Ω, absolutely continuous with respect to $|e^D(u)| + (\operatorname{div} u)^2$.*

(ii) *$\bar{\Psi}$ is a convex function of u.*

Proof. (i) We begin to show that for every u in $U(\Omega)$, $\bar{\Psi}^-(u)$ and $\bar{\Psi}^+(u)$ are Lipschitz with respect to $\varphi \geqq 0$ in $\mathscr{C}(\bar{\Omega})$. We proceed toward that aim in several steps. For convenience we give them only for $\bar{\Psi}^-(u)$, the changes for $\bar{\Psi}^+(u)$ being obvious.

Step 1. For every $\varphi \in \mathscr{C}(\bar{\Omega})$, $\varphi \geqq 0$,

$$\lim_{n \to +\infty} \left\langle \bar{\Psi}^-(u), \operatorname{Sup}\left(\varphi, \frac{1}{n}\right)\right\rangle = \langle \bar{\Psi}^-(u), \varphi\rangle.$$

Step 2. For every φ_1 and φ_2 in $\mathscr{C}^1(\bar{\Omega})$, φ_1 and $\varphi_2 \geqq 0$, we have

$$|\langle \bar{\Psi}^-(u), \varphi_1\rangle - \langle \bar{\Psi}^-(u), \varphi_2\rangle| \leqq C_2(u)|\varphi_1 - \varphi_2|_\infty.$$

Suppose for now that steps 1 and 2 have been proved, assume that φ_1 and φ_2 have been proved, and take φ_1 and φ_2 in $\mathscr{C}(\bar{\Omega})$, φ_1 and $\varphi_2 \geqq 0$. Define $\varphi_i^n = \operatorname{Sup}(\varphi_i, 1/n)$, for $i = 1, 2$. There exists $\bar{\varphi}_i^n \in \mathscr{C}^1(\bar{\Omega})$, $|\bar{\varphi}_i^n - \varphi_i^n| \leqq 1/2n$. Then $\theta_i^n = \bar{\varphi}_i^n - 1/2n$ verifies

$$0 \leqq \theta_i^n \leqq \varphi_i^n \leqq \theta_i^n + \frac{1}{n},$$

and then by step 1,

$$\begin{aligned} \langle \bar{\Psi}^-(u), \varphi_1\rangle &\leqq \langle \bar{\Psi}^-(u), \varphi_1^n\rangle \\ &\leqq \left\langle \bar{\Psi}^-(u), \theta_1^n + \frac{1}{n}\right\rangle \end{aligned}$$

(because $\langle \bar{\Psi}^-(u), \cdot\rangle$ is obviously increasing with respect to $\varphi \geqq 0$).

---

[1] The definition and results that follow are also valid for $\varphi \in \mathscr{C}(\Omega)$, φ bounded.

[2] It is well known that $LU(\Omega) \subset L^q(\Omega)$, for all $q \in [1, N/N-1]$.

By step 2 we then obtain

$$\langle\bar{\Psi}^-(u),\varphi_1\rangle \leqq \langle\bar{\Psi}^-(u),\theta_2^n\rangle + \left|\theta_1^n+\frac{1}{n}-\theta_2^n\right|_\infty$$
$$\leqq \langle\bar{\Psi}^-(u),\varphi_2^n\rangle + |\varphi_1^n-\varphi_2^n|_\infty + \frac{2}{n}$$
$$\leqq \langle\bar{\Psi}^-(u),\varphi_2\rangle + \frac{C_2(u)}{2n} + \frac{2}{n} + |\varphi_1-\varphi_2|_\infty.$$

n being arbitary, this proves the desired result, by changing φ_1 in φ_2.

Proof of step 1. Take $\delta>0$ and $\varphi\geqq 0$ in $\mathscr{C}(\Omega)$, $\delta<|\varphi|_\infty/2$. We easily have $\langle\bar{\Psi}^-(u),\varphi\rangle\leqq\langle\bar{\Psi}^-(u),\sup(\delta,\varphi)\rangle$. Let us take

$$\theta_\delta = \begin{cases} 1 & \text{on } \{x,\varphi(x)\geqq 2\delta\},\\ 0 & \text{on } \{x,\varphi(x)\leqq\delta\},\end{cases}$$

$0\leqq\theta_\delta\leqq 1$, $|\nabla\theta_\delta|_\infty\leqq C_3(\delta)$, $\theta_\delta\in\mathscr{C}^1(\Omega)$, and let $u_\varepsilon^\delta\in LU(\Omega)$,

$$\int\psi\left(\frac{x}{\varepsilon},e(u_\varepsilon^\delta)(x)\right)\varphi\leqq\langle\bar{\Psi}^-(u),\varphi\rangle+\delta,\qquad |u_\varepsilon^\delta-u|_{L^p}\leqq\frac{\delta}{C_3(\delta)}.$$

Now Theorem 3.4 of [21] shows that there exists $\bar{u}_\varepsilon^\delta\in LU(\Omega)$,

$$\left|\int_\Omega|e(\bar{u}_\varepsilon^\delta)|-\int_\Omega|e(u)|\right|\leqq\delta,$$
$$|\operatorname{div}\bar{u}_\varepsilon^\delta-\operatorname{div}u|_2\leqq\delta,$$
$$|\bar{u}_\varepsilon^\delta-u|_{L^p}\leqq\frac{\delta}{C_3(\delta)}.$$

Let us define $w_\varepsilon=u_\varepsilon\theta+\bar{u}_\varepsilon(1-\theta)$. (We drop the dependence on δ.) We have $w_\varepsilon\mapsto u$ in $L^p(\Omega)$, and $e(w_\varepsilon)=e(u_\varepsilon)\theta+e(\bar{u}_\varepsilon)(1-\theta)+(u_\varepsilon-\bar{u}_\varepsilon)\otimes\nabla\theta$. By a technical result of Giaquinta and Modica [14], there exists $z_\varepsilon\in W^{1,p}(\Omega)$ such that

$$\operatorname{div}z_\varepsilon=-\operatorname{tr}((u_\varepsilon-\bar{u}_\varepsilon)\otimes\nabla\theta),$$
$$z_{\varepsilon|\Gamma}=-\frac{1}{\Gamma}\int_\Omega\operatorname{tr}(u_\varepsilon-\bar{u}_\varepsilon)\otimes\nabla\theta,$$
$$|z_\varepsilon|_{W^{1,p}}\leqq C_4|(u_\varepsilon-\bar{u}_\varepsilon)\otimes\nabla\theta|_{L^p(\Omega)}\leqq C_4\delta$$

where the constant C_4 above depends only on N, Ω, and p.

We now define $v_\varepsilon=w_\varepsilon+z_\varepsilon$. We have

$$e(v_\varepsilon)=e(u_\varepsilon)\theta+(1-\theta)e(\bar{u}_\varepsilon)+e^D(z_\varepsilon)+((u_\varepsilon-\bar{u}_\varepsilon)\oplus\nabla\theta)^D$$

and then

$$\operatorname{div}v_\varepsilon=\theta_\delta\operatorname{div}u_\varepsilon+(1-\theta_\delta)\operatorname{div}\bar{u}_\varepsilon$$

so $v_\varepsilon\to v$ in $L^p(\Omega)$ and $v_\varepsilon\in LU(\Omega)$:

$$\int_\Omega|e(v_\varepsilon)-\theta e(u_\varepsilon)-(1-\theta)e(\bar{u}_\varepsilon)|=\int_\Omega|e^D(v_\varepsilon)-\theta e^D(u_\varepsilon)-(1-\theta)e^D(\bar{u}_\varepsilon)|$$
$$\leqq|e(z_\varepsilon)+\nabla\theta\otimes(u_\varepsilon-\bar{u}_\varepsilon)|_{L^1}$$
$$\leqq(C_4+1)\delta\,(\text{meas. }\Omega)^{1-1/p}$$
$$\leqq C_5\delta.$$

And then, using (1.3)

$$|\Psi(x)-\Psi(y)| \leqq C_1|x^D - y^D|,$$

$$\begin{aligned}\int \Psi\left(\frac{x}{\varepsilon}, e(v_\varepsilon)\right) \mathrm{Sup}\,(\varphi,\delta) &\leqq \int \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)\theta + (1-\theta)e(\bar{u}_\varepsilon)\right) \mathrm{Sup}\,(\varphi,\delta)\\ &\quad + C_5(u)\delta(|\varphi|_\infty + \delta)\\ &\leqq \int_{\varphi \geqq \delta} \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)\right)\theta_\delta\varphi + \int_{\varphi \leqq 2\delta} \Psi\left(\frac{x}{\varepsilon}, e(\bar{u}_\varepsilon)\right)\\ &\quad \cdot (1-\theta)\,\mathrm{Sup}\,(\varphi,\delta) + C_5\delta(|\varphi|_\infty + \delta)\\ &\leqq \int \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)\right)(x)\varphi + 2\delta C_1\\ &\quad \cdot \left(\int (|e^D(\bar{u}_\varepsilon)| + |\mathrm{div}\,\bar{u}_\varepsilon|^2) + \mathrm{meas.}\,\Omega\right) + C_5\delta(|\varphi|_\infty + \delta)\end{aligned}$$

and then

$$\langle \bar{\Psi}^-(u), \mathrm{Sup}\,(\varphi,\delta)\rangle \leqq \langle \bar{\Psi}^-(u), \varphi\rangle + 2\delta C_1(\|u\|_{LU(\Omega)} + \delta + \mathrm{meas.}\,\Omega) + C_5\delta(|\varphi|_\infty + \delta).$$

This implies the desired result, δ being arbitrary.

Proof of step 2. We want then to show that for every u in $U(\Omega)$, $\bar{\Psi}^-(u)$ and $\bar{\Psi}^+(u)$ are Lipschitz with respect to φ, when $\varphi \in W^{1,\infty}(\Omega)$, for the norm of $\mathscr{C}(\Omega)$. Toward that aim, let $\delta > 0$ be given, φ_1 and φ_2 in $W^{1,\infty}(\Omega)$, $\varphi_1, \varphi_2 \geqq 0$, and let u_ε be in $LU(\Omega)$, such that $u_\varepsilon \mapsto u$ in $L^p(\Omega)$ and

$$\underline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)\right)(x)\varphi_1(x) \leqq \langle \bar{\Psi}^-(u), \varphi_1\rangle + \delta.$$

By Theorem (3.4) of [21], there exists $\bar{u}_\varepsilon \in LU(\Omega)$, $\bar{u}_\varepsilon \mapsto u$ in $L^p(\Omega)$, $\mathrm{div}\,\bar{u}_\varepsilon \mapsto \mathrm{div}\,u$ in $L^2(\Omega)$ and $\int_\Omega |e^D(\bar{u}_\varepsilon)| \mapsto \int_\Omega |e^D(u)|$. (We will denote it for convenience as the tight convergence in $U(\Omega)$.) We define $\theta = \inf(1, (\varphi_1/(\varphi_2+\delta)))$. It is easy to see that $\theta \in W^{1,\infty}$. Define $w_\varepsilon = u_\varepsilon\theta + \bar{u}_\varepsilon(1-\theta)$. We have $w_\varepsilon \mapsto u$ in $L^p(\Omega)$, and $e(w_\varepsilon) = e(u_\varepsilon)\theta + e(\bar{u}_\varepsilon)(1-\theta) + (u_\varepsilon - \bar{u}_\varepsilon)\otimes\nabla\theta$. By a technical result of Giaquinta and Modica [14], there exists $z_\varepsilon \in W^{1,p}(\Omega)$ such that

$$\begin{aligned}&\mathrm{div}\, z_\varepsilon = -\mathrm{tr}\,((u_\varepsilon - \bar{u}_\varepsilon)\otimes\nabla\theta),\\ &z_{\varepsilon|\Gamma} = -\frac{1}{\Gamma}\int_\Omega \mathrm{tr}\,(u_\varepsilon - \bar{u}_\varepsilon)\otimes\nabla\theta,\\ &|z_\varepsilon|_{W^{1,p}} \leqq C_4|(u_\varepsilon - \bar{u}_\varepsilon)\otimes\nabla\theta|_{L^p(\Omega)}\end{aligned}$$

where the constant above depends only on N, Ω, and p. Since $(u_\varepsilon - \bar{u}_\varepsilon) \mapsto 0$ in L^p, we may choose ε sufficiently small to have

$$|z_\varepsilon|_{W^{1,p}} \leqq C_4\delta.$$

We now define $v_\varepsilon = w_\varepsilon + z_\varepsilon$. We have

$$\begin{aligned}\int_\Omega |e(v_\varepsilon) - \theta e(u_\varepsilon) - (1-\theta)e(\bar{u}_\varepsilon)| &= \int_\Omega |e^D(v_\varepsilon) - \theta e^D(u_\varepsilon) - (1-\theta)e^D(\bar{u}_\varepsilon)|\\ &\leqq |e(z_\varepsilon) + \nabla\theta\otimes(u_\varepsilon - v_\varepsilon)|_{L^1}\\ &\leqq (C_4+1)\delta(\mathrm{meas.}\,\Omega)^{1-1/p}\\ &\leqq C_5\delta.\end{aligned}$$

Then, by (1.3),

$$|\Psi(x)-\Psi(y)| \leqq C_1|x^D-y^D|,$$

$$\int \Psi\left(\frac{x}{\varepsilon}, e(v_\varepsilon)\right)\varphi_2 \leqq \int \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)\theta+(1-\theta)e(\bar{u}_\varepsilon)\right)\varphi_2 + C_5\delta|\varphi_2|.$$

This implies

$$\begin{aligned}\langle\bar{\Psi}^-(u), \varphi_2\rangle &\leqq \underline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, e(v_\varepsilon)\right)\varphi_2 \\ &\leqq \underline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)\right)\theta\varphi_2 + \int \Psi\left(\frac{x}{\varepsilon}, e(\bar{u}_\varepsilon)\right)(1-\theta)\varphi_2 + C_5\delta|\varphi_2|_\infty.\end{aligned}$$

We now remark that $\theta\varphi_2 \leqq \varphi_1$ and obtain

$$\begin{aligned}\langle\bar{\Psi}^-(u), \varphi_2\rangle &\leqq \underline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)\right)\varphi_1 + \overline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, e(\bar{u}_\varepsilon)\right)(1-\theta)\varphi_2 + C_5\delta|\varphi_2| \\ &\leqq \langle\bar{\Psi}^-(u), \varphi_1\rangle + C_1\,\overline{\lim} \int_\Omega (|e^D(\bar{u}_\varepsilon)| + (\operatorname{div}\bar{u}_\varepsilon)^2+1)(|\varphi_2-\varphi_1|_\infty+\delta) \\ &\quad + C_5\delta|\varphi_2|_\infty \\ &\leqq \bar{\Psi}^-(u, \varphi_1) + C_1\int_\Omega (|e^D(u)| + (\operatorname{div} u)^2+1)|\varphi_2-\varphi_1|_\infty + C_6\delta.\end{aligned}$$

Since δ is arbitrary, we obtain that

$$\bar{\Psi}^-(u, \varphi_2) \leqq \bar{\Psi}^-(u, \varphi_1) + C_1\|u\|_{LU(\Omega)}|\varphi_2-\varphi_1|_\infty$$

where $\|u\|_{LU(\Omega)} = \int_\Omega (|e^D(u)| + (\int_\Omega (\operatorname{div} u)^2)^{1/2}$.

This ends the proof of Steps 1 and 2 and thus of the Lipschitz property of $\bar{\Psi}^-$. The analogous is true for $\bar{\Psi}^+$, following the same arguments. Now let $\mathscr{D} = \{\varphi_i\}$ be a countable dense subset in $\mathscr{C}_0^+(\Omega)$. Fix j in $\mathbb{N}$; there exists a subsequence $\sigma_j(\varepsilon)$ such that the corresponding $F_{\sigma_j}(\cdot, \varphi_j)$ are Γ-convergent. (This is a consequence of a Buttazzo compacity result, since L^p possesses a countable base.) By a diagonal process we then obtain that there exists a subsequence $\sigma(\varepsilon)$ such that for every u in $\mathscr{U}(\Omega)$, we have

$$\begin{aligned}\langle\bar{\Psi}^+(u), \varphi_j\rangle &= \langle\bar{\Psi}^-(u), \varphi_j\rangle \\ &= \Gamma_- \lim_{\substack{\varepsilon\to 0 \\ v\mapsto u}} \int \Psi\left(\frac{x}{\sigma(\varepsilon)}, e(v)(x)\right)\varphi_j(x).\end{aligned}$$

Then take u in $U(\Omega)$ and φ in $\mathscr{C}(\bar{\Omega})$, $\varphi \geqq 0$, for all $\varepsilon > 0$; there exists φ_j in $\mathscr{D}$, such that $|\varphi_j - \varphi|_\infty < \varepsilon/C_1(u)$. We then have

$$\begin{aligned}|\langle\bar{\Psi}^-(u), \varphi\rangle - \langle\bar{\Psi}^+(u), \varphi\rangle| &\leqq |\langle\bar{\Psi}^-(u), \varphi\rangle - \langle\bar{\Psi}^-(u), \varphi_j\rangle| + |\langle\bar{\Psi}^+(u), \varphi\rangle - \langle\bar{\Psi}^+(u), \varphi_j\rangle| \\ &\leqq 2C_1(u)|\varphi-\varphi_j|_\infty \\ &\leqq 2\varepsilon.\end{aligned}$$

With ε being arbitrary, this implies that for the subsequence $\sigma(\varepsilon)$, $\langle\bar{\Psi}^-(u), \varphi\rangle = \langle\bar{\Psi}^+(u), \varphi\rangle$, for every φ in $\mathscr{C}(\bar{\Omega})$, $\varphi \geqq 0$. We denote the common value $\bar{\Psi}^-(u) = \bar{\Psi}^+(u)$ by $\bar{\Psi}(u)$. Now let us show that $\bar{\Psi}(u)$ may be extended as a bounded measure on Ω.

For the first time we show the additivity of $\bar{\Psi}(u)$ on $W^{1,\infty}(\Omega)$. Let $\varphi_i \geqq 0$ $(i=1,2)$ be in $W^{1,\infty}(\Omega)$, and u_ε be in $LU(\Omega)$, such that $u_\varepsilon \mapsto u$ in $L^1(\Omega)$, $\int \Psi(x/\varepsilon, e(u_\varepsilon))(\varphi_1+\varphi_2)(x) \to \langle \bar{\Psi}(u), \varphi_1+\varphi_2\rangle$. By the definition of Γ^- limits,

$$\langle \bar{\Psi}(u), \varphi_i\rangle \leqq \underline{\lim}_{\varepsilon\mapsto 0} \int \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)\right)\varphi_i \quad \text{for } i=1,2,$$

and then $\sum_i \langle \bar{\Psi}(u), \varphi_i\rangle \leqq \langle \bar{\Psi}(u), \varphi_1+\varphi_2\rangle$. For the reverse inequality, with $\delta>0$ being given, let $u_\varepsilon^{i\delta}$ be in $LU(\Omega)$, such that $u_\varepsilon^{i\delta}$ tends to u in $L^1(\Omega)$ and $\int \Psi(x/\varepsilon, e(u_\varepsilon^{i\delta}))\varphi_i \leqq \langle \bar{\Psi}(u), \varphi_i\rangle + \delta$, for $i=1, 2$, and $\bar{u}_\varepsilon$ as in the proof of the Lipschitz property of $\bar{\Psi}$, and define

$$u_\varepsilon = \sum_{i=1,2} \frac{\varphi_i u_\varepsilon^{i\delta}}{\sum_i \varphi_i + \delta} + \frac{\delta \bar{u}_\varepsilon}{\sum_i \varphi_i + \delta}.$$

It is easy to see that $u_\varepsilon \in LD(\Omega)$ and that $u_\varepsilon \to u$ in $L^p(\Omega)$. The calculus of $e(u_\varepsilon)$ gives

$$e(u_\varepsilon) = \sum_i \theta_i e(u_\varepsilon^{i\delta}) + \frac{\delta}{\sum_i \varphi_i + \delta} e(\bar{u}_\varepsilon) + (u_\varepsilon^1 - \bar{u}_\varepsilon) \oplus \nabla\theta + (u_\varepsilon^2 - \bar{u}_\varepsilon) \oplus \theta_2$$

where $\theta_i = \varphi_i/(\sum_i \varphi_i + \delta)$ for $i=1,2$.

By the technical result of Giaquinta and Modica [14] already cited, let w_ε be in $W^{1,p}(\Omega)$, such that

$$\begin{aligned}
&\operatorname{div} w_\varepsilon = \operatorname{tr} \lambda_\varepsilon,\\
&w_\varepsilon \cdot n_{|\gamma} = -1/|\Gamma| \int_\Omega \operatorname{tr} \lambda_\varepsilon,\\
&|w_\varepsilon|_{W^{1,p}} \leqq C_4 |(u_\varepsilon^1 - \bar{u}_\varepsilon) \otimes \nabla\theta_1|_{L^p} + |(u_\varepsilon^2 - \bar{u}_\varepsilon) \otimes \theta_2|_{L^p}.
\end{aligned}$$

We may choose ε sufficiently small in order to have the last quantity on the right above less than δ. We then define $v_\varepsilon = u_\varepsilon + w_\varepsilon$. We have $v_\varepsilon \mapsto u$ in $L^p(\Omega)$, and

$$\begin{aligned}
|e(v_\varepsilon) - \theta_1 e(u_\varepsilon^1) - \theta_2 e(u_\varepsilon^2)| &= |e^D(v_\varepsilon) - \theta_1 e^D(u_\varepsilon^1) - \theta_2 e^D(u_\varepsilon^2)|\\
&\leqq |e(w_\varepsilon)| + |(1-\theta_1-\theta_2)e(\bar{u}_\varepsilon)| + |\nabla\theta_1 \otimes (u_\varepsilon^1 - \bar{u}_\varepsilon)| + |\nabla\theta_2 \otimes (u_\varepsilon^2 - \bar{u}_\varepsilon)|.
\end{aligned}$$

We get

$$\begin{aligned}
\langle \bar{\Psi}(u), \varphi_1+\varphi_2\rangle &\leqq \underline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, \theta_1 e(u_\varepsilon^1) + \theta_2 e(u_\varepsilon^2)\right)(\varphi_1+\varphi_2) + (C_2+2)\delta|\varphi_1+\varphi_2+\delta|_\infty\\
&\quad + \left(\int_\Omega |e^D(\bar{u}_\varepsilon)| + (\operatorname{div} \bar{u}_\varepsilon)^2\right)\delta\\
&\leqq \underline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, \theta_1 e(u_\varepsilon^1)\right)(\varphi_1+\varphi_2)\\
&\quad + \overline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, \theta_2 e(u_\varepsilon^2)\right)(\varphi_1+\varphi_2) + C_6\delta\\
&\leqq \langle \bar{\Psi}(u), \varphi_1\rangle + \langle \bar{\Psi}(u), \varphi_2\rangle + C_6\delta.
\end{aligned}$$

With δ being arbitrary, this ends the proof of additivity for $\bar{\Psi}(u)$ on $W^{1,\infty}(\Omega) \cap \mathscr{C}^+(\Omega)$, and then on $\mathscr{C}^+(\bar{\Omega})$, by density, since $\bar{\Psi}(u)$ is Lipschitz with respect to φ for the usual norm of $\mathscr{C}(\bar{\Omega})$.

We now extend $\bar{\Psi}(u)$ to $\mathscr{C}(\Omega)$ by setting

$$\langle \bar{\Psi}(u), \varphi\rangle = \langle \bar{\Psi}(u), \varphi_1\rangle - \langle \bar{\Psi}(u), \varphi_2\rangle \tag{1.10}$$

where $\varphi = \varphi_1 - \varphi_2$ is every decomposition of φ in the difference of two nonnegative functions in $\mathscr{C}^+(\bar{\Omega})$. That this does not depend on the decomposition of φ is a consequence of $\bar{\Psi}(u)$'s additivity. Let us now show that $\bar{\Psi}(u)$ is absolutely continuous with respect to $|e^D(u)| + (\operatorname{div} u)^2$. For that purpose, we use a result of approximation in [21]. For u in $U(\Omega)$, there exists u_ε in $LU(\Omega)$, u_ε tends to u in $L^1(\Omega)$, $|e^D(u_\varepsilon)| \mapsto |e^D(u)|$ tightly on Ω, and $\operatorname{div} u_\varepsilon \mapsto \operatorname{div} u$ in $L^2(\Omega)$. By definition of $\bar{\Psi}(u)$,

$$\begin{aligned} \langle \bar{\Psi}(u), \varphi \rangle &\leqq \underline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)\right)\varphi \\ &\leqq \underline{\lim} \int c_1(|e^D(u_\varepsilon)| + (\operatorname{div} u_\varepsilon)^2 + 1)\varphi \qquad (1.11) \\ &= \int c_1(|e^D(u)| + (\operatorname{div} u)^2 + 1)\varphi. \end{aligned}$$

(ii) It is easy to show that $\bar{\Psi}$ is a convex function with respect to u. Take φ in $\mathscr{C}_0(\Omega)$, $\varphi \geqq 0$, a real t in $[0,1]$, and functions u_ε^i in $LU(\Omega)$, with u_ε^i tending to u^i in $L^p(\Omega)$ $(i = 1, 2)$ such that $\int \Psi(x/\varepsilon, e(u_\varepsilon^i))\varphi \mapsto \langle \bar{\Psi}(u^i), \varphi \rangle$. By the convergence of $tu_\varepsilon^1 + (1-t)u_\varepsilon^2$ toward u in $L^1(\Omega)$ and the properties of the Γ-limit,

$$\begin{aligned} \langle \bar{\Psi}(tu^1 + (1-t)u^2), \varphi \rangle &\leqq \underline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, te(u_\varepsilon^1) + (1-t)e(u_\varepsilon^2)\right)\varphi \\ &\leqq \overline{\lim} \int t\Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon^1)\right)\varphi + \overline{\lim} \int (1-t)\Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon^2)\right)\varphi \\ &= t\langle \bar{\Psi}(u^1), \varphi \rangle + (1-t)\langle \bar{\Psi}(u^2), \varphi \rangle. \end{aligned}$$

PROPOSITION 1.1. *Let u be in $LU(\Omega)$. There exists u_ε in $LU(\Omega)$, u_ε tends to u tightly in $LU(\Omega)$, $u_\varepsilon = u_{|\Gamma}$, and*

$$\int_\Omega \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)\right) \to \int_\Omega \bar{\Psi}(u).$$

COROLLARY. *For u in $LU(\Omega)$, there exists u_ε in $LU(\Omega)$, $u_\varepsilon \mapsto u$ tightly in $LU(\Omega)$, $u_\varepsilon = u_{|\Gamma}$ and $\Psi(x/\varepsilon, e(u_\varepsilon)) \mapsto \bar{\Psi}(u)$ tightly on Ω.*

Proof. Assume that u belongs to $LU(\Omega)$. We are going to show that for every $\delta > 0$ there exists $u_{\varepsilon,\delta}$ in $LU(\Omega)$, such that

$$|u_{\varepsilon,\delta} - u|_1 < \delta, \qquad (1.12)$$

$$\int_\Omega \Psi\left(\frac{x}{\varepsilon}, e(u_{\varepsilon,\delta})\right) \leqq \int_\Omega \bar{\Psi}(u) + \delta, \qquad (1.13)$$

$$u_{\varepsilon,\delta} = u_{|\Gamma}. \qquad (1.14)$$

Let us show (1.12)-(1.14). For $\delta > 0$ given, let κ_δ be a compact subset in Ω such that $\int_{\Omega \setminus \kappa_\delta} |e(u)| < \delta$ and a function φ_δ in $\mathscr{C}_0(\Omega)$, $\varphi_\delta = 1$ on κ_δ, $0 \leqq \varphi_\delta \leqq 1$. By definition of $\int_\Omega \bar{\Psi}(u)$, there exists $v_{\varepsilon,\delta}$ in $LU(\Omega)$ such that

$$\int_\Omega |v_{\varepsilon,\delta} - u|_p < \frac{\delta}{|\nabla \varphi_\delta|_\infty},$$

$$\int_\Omega \Psi\left(\frac{x}{\varepsilon}, e(v_{\varepsilon\delta})\right) \leqq \int_\Omega \bar{\Psi}(u) + \delta.$$

This last inequality and (1.2) imply that $v_{\varepsilon\delta}$ is bounded in $LU(\Omega)$ and then is weakly convergent toward u in $U(\Omega)$. We define $z_{\varepsilon,\delta}=\varphi_\delta v_{\varepsilon,\delta}+(1-\varphi_\delta)u$. By the result of Giaquinta and Modica [14], let $w_{\varepsilon,\delta}$ be in $W^{1,p}$ for some $p\in]1, N/N-1[$

$$\operatorname{div}\tilde{w}_{\varepsilon\delta}=\nabla\varphi_\delta\cdot(v_{\varepsilon\delta}-u)-\frac{1}{|\Omega|}\int_\Omega(\nabla\varphi_\delta\cdot(v_{\varepsilon\delta}-u)),$$

$$\tilde{w}_{\varepsilon\delta}=0_{|\Gamma},$$

$$|\tilde{w}_{\varepsilon\delta}|_{W^{1,p}}\leqq 2c|(\nabla\varphi_\delta)(v_{\varepsilon\delta}-u)|_p$$

and define $u_{\varepsilon\delta}=v_{\varepsilon\delta}-w_{\varepsilon\delta}$; $\operatorname{div}u_\varepsilon=\varphi_\delta\operatorname{div}v_{\varepsilon\delta}+(1-\varphi_\delta)\operatorname{div}u+(1/|\Omega|)\int_\Omega\nabla\varphi_\delta\cdot(v_{\varepsilon\delta}-u)$ tends to $\operatorname{div}u$ in $L^2(\Omega)$. Let us calculate $e(u_\varepsilon)$:

$$\begin{aligned}e(u_\varepsilon)&=\varphi_\delta e(v_{\varepsilon\delta})+(1-\varphi_\delta)e(u)+(\nabla\varphi_\delta)\otimes(v_{\varepsilon\delta}-u)-e(w_{\varepsilon\delta})\\&=\varphi_\delta e(v_{\varepsilon\delta})+(1-\varphi_\delta)e(u)+[\nabla\varphi_\delta\otimes(v_{\varepsilon\delta}-u)]^D-e^D(w_{\varepsilon\delta})\\&\quad+\frac{1}{|\Omega|}\left(\int_\Omega\nabla\varphi_\delta\cdot(v_{\varepsilon\delta}-u)\right)\mathrm{Id}.\end{aligned}$$

Using the convexity of Ψ and (1.2) and (1.3), we get

$$\begin{aligned}\psi\left(\frac{x}{\varepsilon},e(u_\varepsilon)\right)&\leqq\int_\Omega\varphi_\delta\Psi\left(\frac{x}{\varepsilon},e(v_{\varepsilon\delta})\right)+\int_\Omega(1-\varphi_\delta)\Psi\left(\frac{x}{\varepsilon},e(u)\right)\\&\quad+C_1\left|\int_\Omega|e^D(\tilde{w}_{\varepsilon\delta})|+\int_\Omega|\nabla\varphi_\delta\otimes(v_{\varepsilon\delta}-u)^D|\right|\\&\quad+C_1\left[\frac{1}{N|\Omega|}\int_\Omega\nabla\varphi_\delta\cdot(v_{\varepsilon\delta}-u)\right]^2|\Omega|\\&\leqq\int_\Omega\bar{\Psi}(u)+\delta+\int_{\Omega\setminus\kappa_\delta}|e(u)|+C_1\left[2C\delta+\delta+\frac{1}{N^2}\delta^2|\Omega|\right]\\&\leqq\int_\Omega\bar{\Psi}(u)+C_2\delta,\end{aligned}$$

which ends the proof of Proposition 1.1.

Proof of the corollary. Let u be in $LU(\Omega)$ and u_ε as in Proposition 1.1. Since u_ε is tightly convergent toward u, $\Psi(x/\varepsilon, e(u_\varepsilon))$ is (due to (1.2)) a bounded sequence in $L^1(\Omega)$, we may then extract from it a subsequence, still denoted u_ε such that $\Psi(x/\varepsilon, e(u_\varepsilon))\mapsto\nu$ vaguely. By the definition of the Γ-limit $\bar{\Psi}(u)$, we have for every φ in $\mathscr{C}^0(\Omega)$, $\varphi\geqslant 0$,

$$\begin{aligned}\langle\bar{\Psi}(u),\varphi\rangle&\leqslant\underline{\lim}\int\Psi\left(\frac{x}{\varepsilon},e(u_\varepsilon)(x)\right)\varphi(x)\,dx\\&=\int\nu\varphi,\end{aligned}$$

and then $\bar{\Psi}(u)\leqslant\nu$. On the other hand, by lower semicontinuity we have an open set

$$\begin{aligned}\int\nu&\leqslant\underline{\lim}\int\Psi\left(\frac{x}{\varepsilon},e(u_\varepsilon)(x)\right)\varphi(x)\,dx\\&=\int\bar{\Psi}(u)\end{aligned}$$

(by definition of u_ε). We finally obtain $\nu = \bar{\Psi}(u)$; thus every subsequence of $\Psi(x/\varepsilon, e(u_\varepsilon)(x))$ converges to $\nu = \bar{\Psi}(u)$ and then all the sequence tends to it.

Furthermore, we will need the following locally Lipschitz result for $\bar{\Psi}$. Let $\mathcal{O}$ be an open set in Ω. We denote for u in $LU(\Omega)$ by $\|u\|_\partial$ the quantity $\|u\|_\mathcal{O} = \int_\mathcal{O} |e^D(u)| + (\int_\mathcal{O} |\text{div } u|^2)^{1/2}$. We then have the following result.

PROPOSITION 1.2. *For every $R>0$ and for every $(u, v) \in LU(\Omega)$ subject to $\int_\Omega |e^D(v)| + (\text{div } v)^2$ less than R, for every $\mathcal{O}$ an open set in Ω, we have*

$$\left| \int_\mathcal{O} (\bar{\Psi}(u) - \bar{\Psi}(v)) \right| \leqq C_1(3R+4)\|u-v\|_\mathcal{O}. \tag{1.15}$$

Proof.

- Suppose first that $\|u-v\|_\mathcal{O} > 1$. Then

$$\begin{aligned} \left| \int_\mathcal{O} (\bar{\Psi}(u) - \bar{\Psi}(v)) \right| &\leqq \int_\mathcal{O} (\bar{\Psi}(u) + \bar{\Psi}(v)) \\ &\leqq 2C_1(R+1) \\ &\leqq \|u-v\|_\mathcal{O} 2C_1(R+1). \end{aligned}$$

- Assume now that $0 < \|u-v\|_\mathcal{O} \leqq 1$ and define $w = u + (v-u)/(\|u-v\|_\mathcal{O})$; we have $v = u + \|u-v\|_\mathcal{O}(w-u)$, and by the convexity of $\bar{\Psi}$,

$$\begin{aligned} \bar{\Psi}(v) - \bar{\Psi}(u) &\leqq \|u-v\|_\mathcal{O}(\bar{\Psi}(w) - \bar{\Psi}(u)) \\ &\leqq \|u-v\|_\mathcal{O}(|\bar{\Psi}(w)| + |\bar{\Psi}(u)|) \\ &\leqq \|u-v\|_\mathcal{O}(3C_1R+4) \end{aligned}$$

$$\left[\text{because } \int_\mathcal{O} \bar{\Psi}(w) \leqq C_1 \left(\int_\mathcal{O} |e^D(w)| + \int |\text{div } w|^2 \right) \right.$$

$$\leqq C_1 \left(\int_\mathcal{O} |e^D(u)| + 1 + 2((\text{div } u)^2 + 1) \right)$$

$$\left. \leqq C_1(2R+3) \right].$$

By changing u and v we get the desired result.

- We finally deal with the case where $\|u-v\|_\mathcal{O} = 0$, and show that $\int_\mathcal{O} (\bar{\Psi}(u) - \bar{\Psi}(v)) = 0$. Toward that aim, by the definition of $\int_\mathcal{O} \bar{\Psi}(v)$, let φ_δ be in $\mathcal{C}_0(\Omega)$, $0 \leqq \varphi_\delta \leqq 1$, with a compact support in Ω, such that $\int_\mathcal{O} \bar{\Psi}(v)(1-\varphi_\delta) < \delta$. Also, let u_ε be in $LU(\Omega)$, subject to $u_\varepsilon \to u$ in $L^p(\Omega)$, and $\int \Psi(x/\varepsilon, e(u_\varepsilon)(x))\varphi_\delta \to \int \bar{\Psi}(u)\varphi_\delta$. The sequence $u_\varepsilon + v - u$ converges toward v, and then

$$\begin{aligned} \langle \bar{\Psi}(v), \varphi_\delta \rangle &\leqq \underline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon + v - u)\right)(x)\varphi_\delta \\ &\leqq \underline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)(x)\right)\varphi_\delta(x)\, dx + C_1 \int (|e^D(v-u)| + |\text{div } (v-u)|^2)\varphi_\delta \\ &\leqq \int_\mathcal{O} \bar{\Psi}(u)\varphi_\delta + 0 \\ &\leqq \int_\mathcal{O} \bar{\Psi}(u). \end{aligned} \tag{1.16}$$

with δ being arbitrary, we obtain the desired conclusion by changing u and v.

We conclude this section by giving two useful results. One is a lemma of "commutativity" of $\bar{\Psi}$ with translation, which is due to the periodicity of Ψ. The other may be considered as a Jensen's inequality for the measure $\bar{\Psi}(u)$.

LEMMA 1.1. *Assume that u belongs to $U(\Omega)$, φ is in $\mathscr{C}_0(\Omega)$, and $y \in \mathbb{R}^N$ is such that $|y| < d(\mathrm{Supp}\ \varphi, \partial\Omega)$. Then*

$$\langle \bar{\Psi}(u)(\cdot - y), \varphi\rangle = \langle \bar{\Psi}(u), \varphi(\cdot + y)\rangle. \quad [3] \tag{1.17}$$

Proof. We will obtain (1.17) by showing the inequality

$$\langle \bar{\Psi}(u)(\cdot - y), \varphi\rangle \leqq \langle \bar{\Psi}(u), \varphi(\cdot + y)\rangle$$

for φ in $\mathscr{C}_0(\Omega) \cap W^{1,\infty}(\Omega)$, $\varphi \geqq 0$, and y, $|y| < d(\mathrm{Supp}\ \varphi, \partial\Omega)$; the reverse is obtained by considering the functions $u(\cdot + y)$ and $\varphi(\cdot - y)$ in place of u and φ. So let φ and y be as above. The function $\varphi(\cdot + y)$ is nonnegative in $\mathscr{C}_0(\Omega)$. Let us then take u_ε in $LU(\Omega)$, u_ε tending to u in L^1 strongly, and $\int \Psi(x/\varepsilon, e(u_\varepsilon)(x))\varphi(x+y)\,dx \to \langle \bar{\Psi}(u), \varphi(\cdot + y)\rangle$. Let $y_\varepsilon^i = [y_i/\varepsilon]$.[4] We easily have that $y_\varepsilon = (y_\varepsilon)_i \in \mathbb{Z}^N$, $\varepsilon \vec{y}_\varepsilon \mapsto \vec{y}$, $|\varepsilon \vec{y}_\varepsilon| < d(\mathrm{Supp}^t \cdot \varphi, \partial\Omega)$, and we let $\bar{u}_\varepsilon$ be in $LU(\Omega)$, $\bar{u}_\varepsilon \to u$ in $LU(\Omega)$ tightly, and $\theta_\varepsilon = \varphi(\cdot + y_1)/(\delta + \varphi(\cdot + \varepsilon y_\varepsilon))$. For ε sufficiently small, θ_ε is less than 1. We define $v_\varepsilon = \theta_\varepsilon u_\varepsilon + (1 - \theta_\varepsilon) v_\varepsilon$ and $\tilde{u}_\varepsilon = v_\varepsilon + w_\varepsilon$ in a way we have already used twice:

$$\begin{aligned}
&\operatorname{div} w_\varepsilon = -\operatorname{tr}(u_\varepsilon - v_\varepsilon) \otimes \nabla \theta_\varepsilon, \\
&w_{\varepsilon|\Gamma} = -\frac{1}{|\Gamma|} \int \operatorname{tr}(u_\varepsilon - v_\varepsilon) \otimes \nabla \theta_\varepsilon, \\
&|w_\varepsilon|_{W^{1,p}} \leqq C |\operatorname{tr}(u_\varepsilon - v_\varepsilon) \otimes \nabla \theta_\varepsilon|_{L^p}.
\end{aligned}$$

It is easy to see that $|\nabla \theta_\varepsilon|_\infty$ is bounded independently with respect to ε, and then we may choose ε sufficiently small to obtain $|w_\varepsilon|_{W^{1,p}} \leqq \delta$. By definition of Γ-limits, we may write, since $\tilde{u}_\varepsilon \to u$ in L^p and $\tilde{u}_\varepsilon(\cdot - \varepsilon y_\varepsilon) \mapsto u(\cdot - y)$:

$$\begin{aligned}
\langle \bar{\Psi}(u(\cdot - y)), \varphi\rangle &\leqq \underline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, e(\tilde{u}_\varepsilon)(x - \varepsilon y_\varepsilon)(\varphi(x))\right) \\
&\leqq \underline{\lim} \int \Psi\left(\frac{x + \varepsilon y_\varepsilon}{\varepsilon}, e(\tilde{u}_\varepsilon)(x)(\varphi(x) + \varepsilon y_\varepsilon)\right) \\
&\leqq \underline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)(x)\theta_\varepsilon(x)(\varphi(x) + \varepsilon y_\varepsilon)\right) \\
&\quad + \overline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, e(\bar{u}_\varepsilon)(x)(1 - \theta_\varepsilon)(x)\varphi(x + \varepsilon y_\varepsilon)\right) \\
&\quad + \overline{\lim}\, C_1 \int (|e^D(w_\varepsilon)| + |\nabla \theta_\varepsilon \otimes (u_\varepsilon - \bar{u}_\varepsilon)|)\varphi(x + \varepsilon y_\varepsilon) \\
&\leqq \underline{\lim} \int \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)(x)\right)\varphi(x + y) + C_1\delta \\
&\quad + C_1 \int (|e^D(u)| + (\operatorname{div} u)^2) \sup_x |\varphi(x + \varepsilon y_\varepsilon) - \varphi(x + y)| \\
&\leqq 2C_1\delta + \langle \bar{\Psi}(u), \varphi(\cdot - y)\rangle.
\end{aligned}$$

With δ being arbitrary, we have the result of the lemma.

[3] For a measure ν we define its shifted $\nu(\cdot + y)$ by $\langle \nu(\cdot + y), \varphi\rangle = \langle \nu(\cdot), \varphi(\cdot - y)\rangle$ for all φ in $\mathscr{C}_0(\Omega)$, $y < d(\mathrm{Supp}\ \varphi, \partial\Omega)$.

[4] For a real ξ, $[\xi]$ denotes the entire part of ξ.

We now give a result that we will need in proving the fundamental result of Theorem 2.1 in the next section.

PROPOSITION 1.3. *Let ρ be in $\mathscr{C}_0^\infty(\mathscr{R}^N)$, $\rho \geqq 0$, $\int \rho = 1$, and φ in $\mathscr{C}_0(\Omega)$, $\varphi \geqq 0$, such that*

$$\mathrm{Supp}^t \cdot \varphi + \mathrm{Supp}^t \cdot \rho \subset \Omega.$$

Then for all u in $U(\Omega)$,

$$\langle \bar{\Psi}(u * \rho), \varphi \rangle \leqq \langle \rho * \bar{\Psi}(u), \varphi \rangle.$$

Proof. With $\delta > 0$ being given, let us define $(\Omega_i^\delta)_i$ a partition of Ω into disjointed universally measurable sets, such that meas. $(\Omega_i^\delta) < \delta$. Using the Mean Value Theorem on Ω_i^δ, we are allowed to choose y_i^δ in Ω_i^δ, such that $\rho(y_i^\delta) = 1/|\Omega_i^\delta| \int_{\Omega_i^\delta} \rho(x)\, dx$, so the equality $\sum_i |\Omega_i^\delta| \rho(y_i^\delta) = 1$ holds. It is well known that the function $u^\delta = \sum_i |\Omega_i^\delta| \rho(y_i^\delta) u(\cdot - y_i^\delta)$ is strongly convergent towards $\rho * u$ in $L^1(\Omega)$, so by lower semicontinuity of the Γ-limit, and using the convexity of $\bar{\Psi}$, we have

$$\begin{aligned} \langle \bar{\Psi}(\rho * u), \varphi \rangle &\leqq \varliminf_{\delta \mapsto 0} \langle \bar{\Psi}(u^\delta), \varphi \rangle \\ &\leqq \varliminf_{\delta \mapsto 0} \sum_i |\Omega_i^\delta| \rho(y_i^\delta) \langle \bar{\Psi}(u(\cdot - y_i^\delta)), \varphi \rangle. \end{aligned}$$

We now apply Lemma 1.1 for $y = y_i^\delta$, and use the uniform convergence of $\sum_i |\Omega_i^\delta| \rho(y_i^\delta) \varphi(\cdot + y_i^\delta)$ toward $\check{\rho} * \varphi$, to obtain

$$\begin{aligned} \langle \bar{\Psi}(\rho * u), \varphi \rangle &\leqq \varliminf_{\delta \mapsto 0} \sum_i |\Omega_i^\delta| \rho(y_i^\delta) \varphi(\cdot + y_i^\delta) \\ &= \varliminf_{\delta \mapsto 0} \left\langle \bar{\Psi}(u), \sum_i |\Omega_i^\delta| \rho(y_i^\delta) \varphi(\cdot + y_i^\delta) \right\rangle \\ &= \langle \bar{\Psi}(u), \check{\rho} * \varphi \rangle. \end{aligned}$$

With the left-hand side of the inequality above being nothing but $\langle \rho * \bar{\Psi}(u), \varphi \rangle$, this ends the proof of (i).

2. Integral representation of $\bar{\Psi}(u)$. In this section we want to show the existence of a function Ψ_{hom}, defined on E, such that for all φ in $\mathscr{C}_0(\Omega)$ we have

$$\langle \bar{\Psi}(u), \varphi \rangle = \int \Psi_{\mathrm{hom}}(e(u)(x)) \varphi(x)\, dx$$

when u is sufficiently regular (C^1, for example). The convexity of Ψ_{hom} and its behavior at infinity will permit us, using the theory of convex function of a measure (cf. [11], [12]), to define the measure $\Psi_{\mathrm{hom}}(e(u))$ when u belongs to $U(\Omega)$. The central result of this section will be that

$$\Psi_{\mathrm{hom}}(e(u)) = \bar{\Psi}(u)$$

for every u in $U(\Omega)$.

The existence of the punctual function Ψ_{hom} is a consequence of Proposition 2.1 below.

PROPOSITION 2.1. *There exists a function h on $\mathbb{R}^N \times E$, measurable with respect to x and uniformly locally Lipschitz with respect to the second variable, such that*

$$\langle \bar{\Psi}(u), \varphi \rangle = \int h(x, e(u)(x)) \varphi(x)\, dx \tag{2.1}$$

for every u in $LU(\Omega)$.

Proof. This proof borrows ideas from Lemma 4.1 of [4]. Inequality (1.11) implies, for every $u \in LU(\Omega)$, that $\bar{\Psi}(u)$ is absolutely continuous with respect to the Lebesgue measure. We then denote by h_u the function in $L^1(\Omega)$ defined by

$$\langle \bar{\Psi}(u), \varphi \rangle = \int h_u(x)(\varphi(x))\, dx.$$

Let $\mathcal{P}$ be the set of functions of the form $\xi \cdot x$ where ξ belongs to E, with coordinates in $\mathbb{Q}$. There exists a subset M in Ω, such that meas. $(\Omega - M) = 0$, which verifies the following. If $x \in M$ and $u \in \mathcal{P}$, we define

$$h(x, \xi) = \lim_{\delta \mapsto 0} \frac{1}{|B(x, \delta)|} \int_{B(x,\delta)} h_u(x)\, dx$$

where $u = \xi \cdot x \cdot h$ may be extended to E by continuity. It is clear that $h(x, \cdot)$ is a convex function. Moreover, by (1.11) we have

$$\begin{aligned} |h(x, \xi)| &\leqq \lim_{\delta \mapsto 0} \frac{1}{|B(x, \delta)|} \int_{B(x,\delta)} \bar{\Psi}(u) \\ &\leqq C_1(|e^D(u)| + (\operatorname{div} u)^2 + 1) \\ &\leqq C_1(|\xi^D| + (\operatorname{tr} \xi)^2 + 1). \end{aligned}$$

Then, as in the proof of Proposition 1.2, the functional $\int_A h(x, e(u)(x))\, dx$ verifies the local Lipschitz Property (1.15), where $\bar{\Psi}(u)$ and $\bar{\Psi}(v)$ are replaced by $h(\cdot, e(u)(\cdot))$ and $h(\cdot, e(v)(\cdot))$. Let us then take $u \in LU(\Omega)$, $\delta > 0$, and (A_k) some open disjoint subsets subject to $\cup \overline{A_k} = \Omega$ and meas. $(\partial A_k) = 0$, $(\xi_k)_k \in E$ with rational coordinates, such that

$$\int_\Omega \left| e^D(u) - \sum_k \xi_k^D \chi_{A_k} \right| + \left(\int_\Omega \left[\operatorname{div} u - \sum_k (\operatorname{tr} \xi_k) \chi_{A_k} \right]^2 \right)^{1/2} < \delta.$$

Let us note that we may choose the ξ_k to have in addition

$$\sum_k (|\xi_k^D| + (\operatorname{tr} \xi_k)^2)(\text{meas. } A_k) \leqq \int_\Omega (|e^D(u)| + |\operatorname{div} u|^2) = R.$$

We then obtain

$$\begin{aligned} \left| \int_\Omega h(x, e(u)) - \sum \int_{A_k} h(x, \xi_k) \right| &\leqq \sum \left| \int_{A_k} h(x, e(u)) - h(x, \xi_k) \right| \\ &\leqq \left(\sum_k \int_{A_k} |e^D(u) - \sum \xi_k^D \chi_{A_k}| \right. \\ &\qquad \left. + \left(\int_\Omega \left(\operatorname{div} u - \sum (\operatorname{tr} \xi_k) \chi_{A_k} \right)^2 \right)^{1/2} \right) C_1(3R + 4)\delta. \end{aligned}$$

Now we denote by v_k a function in $LU(\Omega)$, such that $v_k(x)|_{A_k} = \xi_k \cdot x|_{A_k}$, which verifies $\int_\Omega |e^D(v_k)| + |\operatorname{div} v_k|^2 \leqq R$. By Proposition 1.2 for $\bar{\Psi}$ and by using meas. $\partial A_k = 0$, we may write

$$\left| \int_{A_k} \bar{\Psi}(u) - \bar{\Psi}(v_k) \right| \leqq C_1(3R + 4) \| u - v_k \|_{A_k}.$$

Furthermore, by summing on k:

$$\left| \int_\Omega \bar{\Psi}(u) - \sum_k \int_{A_k} \bar{\Psi}(v_k) \right| \leqq C_1(3R + 4)\delta$$

using the fact that $\int_{A_k} \bar{\Psi}(v_k) = \int_{A_k} h(x, \xi_k)$, we have finally obtained that

$$\left| \int_\Omega \bar{\Psi}(u) - \int_\Omega h(x, e(u)(x)) \right| \leqq C_1(3R+4)2\delta.$$

with δ being arbitrary, this ends the proof of Proposition 2.1.

PROPOSITION 2.2. *The function h defined in Proposition 2.1 does not depend on $x \in \mathbb{R}^N$. In other words, there exists a function Ψ_{hom}, defined on E such that*

$$\langle \bar{\Psi}(u), \varphi \rangle = \int \Psi_{\text{hom}}(e(u)(x))\varphi(x)\, dx \tag{2.2}$$

for every u in $LU(\Omega)$.

Proof. It suffices to show that h is locally constant with respect to the first variable. Let φ be in $\mathscr{C}_0(\Omega)$, and y, $|y| < d(\text{Supp } \varphi, \partial\Omega)$. By using Proposition 2.1 and Lemma 1.1 we get

$$\begin{aligned} \int h(x+y, e(u)(x))\varphi(x)\, dx &= \int h(x, e(u_{-y})(x))\varphi_{-y}(x)\, dx \\ &= \langle \bar{\Psi}(u_{-y}), \varphi_{-y} \rangle \\ &= \langle \bar{\Psi}(u), \varphi_{-y+y} \rangle \\ &= \langle \bar{\Psi}(u), \varphi \rangle. \end{aligned}$$

□

Now we want to show that the integral representation formula of $\bar{\Psi}$ may be extended to the u's in $U(\Omega)$. In other words, we want to show Theorem 2.1.

THEOREM 2.1. *Assume that u belongs to $U(\Omega)$; then for every φ in $\mathscr{C}_0(\Omega)$,*

$$\langle \bar{\Psi}(u), \varphi \rangle = \langle \Psi_{\text{hom}}(e(u)), \varphi \rangle \tag{2.3}$$

where $\Psi_{\text{hom}}(e(u))$ is taken as function of a measure (cf. [11], [12]).

Proof. To show (2.3) let us first remark that

$$\Psi_{\text{hom}}(e(u)) \geqq \bar{\Psi}(u).$$

Indeed, by an approximation result that we have used already [21], there exists u_n in $LU(\Omega)$, u_n tends to u in $L^1(\Omega)$, and $\Psi_{\text{hom}}(e(u_n)) \to \Psi_{\text{hom}}(e(u))$ vaguely. Propositions 2.1 and 2.2 give, for φ in $\mathscr{C}_0(\Omega)$,

$$\langle \bar{\Psi}(u_n), \varphi \rangle = \int \Psi_{\text{hom}}(e(u_n))(x)\varphi(x)\, dx,$$

so by lower semicontinuity of the Γ-limits,

$$\begin{aligned} \langle \bar{\Psi}(u), \varphi \rangle &\leqq \underline{\lim} \langle \bar{\Psi}(u_n), \varphi \rangle \\ &= \underline{\lim} \int \Psi_{\text{hom}}(e(u_n)(x))\varphi(x)\, dx \\ &= \int \Psi_{\text{hom}}(e(u)(x))\varphi(x)\, dx. \end{aligned}$$

For the reverse inequality let us take $\varphi \geqq 0$, φ in $\mathscr{C}_0(\Omega)$, $n \in \mathbb{N}$, such that $1/n < d(\text{Supp } \varphi, \partial\Omega)$ and $\rho_n(x) = n^N \rho(nx)$, where ρ is in $\mathscr{C}_0^\infty(\mathbb{R}^N)$, $\int \rho = 1$, $0 \leqq \rho \leqq 1$, $u_n = \rho_n * u$. It is shown in [11] that u_n tends to u and $\int \Psi_{\text{hom}}(e(u_n))\varphi$ tends to $\int_\Omega \Psi_{\text{hom}}(e(u))\varphi$.

Using Proposition 1.3, we then write

$$\begin{aligned}\int \Psi_{\text{hom}}(e(u))\varphi &= \lim_{n\to+\infty} \int \Psi_{\text{hom}}(e(u_n)(x))\varphi(x)\,dx \\ &= \lim_{n\to+\infty} \langle \bar{\Psi}(u_n), \varphi\rangle \\ &= \lim_{n\to+\infty} \langle \bar{\Psi}(\rho_n * u), \varphi\rangle \\ &\leqq \underline{\lim}_{n\to+\infty} \langle \rho_n * \bar{\Psi}(u), \varphi\rangle.\end{aligned}$$

From the classical fact that $\rho_n * \bar{\Psi}(u)$ is vaguely convergent toward $\bar{\Psi}(u)$, the conclusion follows.

Let us remark that it is now possible to extend the approximation result in Proposition 1.2 to the u's in $U(\Omega)$. This may be written as follows.

PROPOSITION 2.3. *Assume that u belongs to $U(\Omega)$. Then there exists u_ε in $LU(\Omega)$, such that*

$$\begin{aligned}&u_\varepsilon \to u \quad \text{in } L^1(\Omega),\\ &u_\varepsilon = u_{|\Gamma},\\ &\operatorname{div} u_\varepsilon \to \operatorname{div} u \quad \text{in } L^2(\Omega) \text{ weakly},\\ &\int \Psi\left(\frac{x}{\varepsilon}, e(u_\varepsilon)\right) \to \int \Psi_{\text{hom}}(e(u)).\end{aligned}$$

Proof. Let $\delta > 0$ be given, and u be in $U(\Omega)$. By an approximation result in [21], there exists v in $LU(\Omega)$ such that $|v-u| < \delta$, $|\operatorname{div} v - \operatorname{div} u|_{L^2(\Omega)} < \delta$, $|\int_\Omega |e^D(u)| - |e^D(v)|| \leqq \delta$, $v = u_{|\Gamma}$, and $|\int \Psi_{\text{hom}}(e(v)) - \Psi_{\text{hom}}(e(u))| < \delta$. By applying Proposition 1.2 to v, we obtain the desired result.

3. An explicit expression of Ψ_{hom}. This section is devoted to the explicit expression of Ψ_{hom} for a particular subsequence. It involves a new method that uses "lower semicontinuity" for convex functions of a measure. The expression we find is

$$\Psi_{\text{hom}}(\xi) = \inf_{u \in LU_{\text{per}}(Y)} \int \Psi(x, \xi + e(u)) \tag{3.1}$$

where $LU_{\text{per}}(Y)$ will be defined later. Formula (3.1) has been given by Tang Qi [19] and Bouchitte [3] for the gradient, but their proofs use, respectively, penalization and dual methods. Before showing (3.1), we need some preliminaries on periodic measures that we did not find in the literature (but they must exist!).

3.1. Preliminaries on periodic measures.

DEFINITION 3.1. We denote by $M^1_{\text{per}}(Y)$ the space of the measures μ in $\mathbb{R}^N$ that verify the property

$$\int_{\mathbb{R}^N} \mu\varphi = \int \mu\varphi_{\vec{j}} \tag{3.2}$$

for every φ continuous with a compact support in Y, $\vec{j} \in \mathbb{Z}^N$, and where $\varphi_{\vec{j}}(x) = \varphi(x + \vec{j})$.

Remark 3.1. Of course, a measure in $M^1_{\text{per}}(Y)$ that is not identically zero cannot be a bounded measure on $\mathbb{R}^N$.

Remark 3.2. This notion coincides with that of usual periodicity when μ is a function. Moreover, $\mu \in M^1_{\text{per}}(Y) \Rightarrow |\mu| \in M^1_{\text{per}}(Y)$.

Remark 3.3. μ is a Y-periodic measure if and only if it is a Y-periodic distribution and also a measure. The if part directly results from Definition 3.1; the only if part is a consequence of Proposition 3.1(i).

PROPOSITION 3.1. (i) *Formula* (3.3) *may be extended to functions* φ *in* $\mathscr{C}(\mathbb{R}^N, X)$, *Y periodic, in the sense that*

$$\int_Y \mu \cdot \varphi = \int_{Y+\vec{j}} \mu \cdot \varphi \quad \textit{for all } \vec{j} \textit{ in } \mathbb{Z}^N.$$

(ii) $$\int_B |\mu| = \int_{B+\vec{j}} |\mu|,$$

for every Borel subset B in $\mathbb{R}^N$, *and for all* $\vec{j}$ *in* $\mathbb{Z}^N$.

Proof. (i) Using Remark 3.2, we may assume that μ is nonnegative. Let us then take a nonnegative function φ in $\mathscr{C}(\mathbb{R}^N, X)$ that is Y-periodic, and a function Ψ in $\mathscr{C}_0(\mathbb{R}^N)$ with compact support in Y, $0 \leqq \Psi \leqq \varphi$, such that $\int_Y \mu\varphi \leqq \int_Y \mu\Psi + \delta$. By applying (1.26) and remarking that $\Psi_{-\vec{j}}$ belongs to $\mathscr{C}_0(Y+\vec{j})$ and $\Psi_{-\vec{j}} \leqq \varphi_{-\vec{j}} = \varphi$, we have

$$\begin{aligned}\int_Y \mu\varphi &\leqq \int_{\mathbb{R}^N} \mu\Psi + \delta \\ &= \int_{\mathbb{R}^N} \mu\Psi_{\vec{j}} + \delta \\ &\leqq \int_{Y+\vec{j}} \mu\varphi + \delta.\end{aligned}$$

The reverse inequality may of course be shown by the same process.

(ii) We begin to show (3.4) when B is a subset of Y and first of all when B is open. We assume once more that μ is nonnegative, $\mathcal{O}$ is an open set in Y, δ is a positive number, and φ is in $\mathscr{C}_0(Y)$, $0 \leqq \varphi \leqq 1_\Omega$ and $\int_{\mathcal{O}} \mu\varphi \geqq \int_{\mathcal{O}} \mu - \delta$. The function $\varphi_{\vec{j}}$ is in $\mathscr{C}_0(\mathcal{O}\backslash\vec{j})$, $0 \leqq \varphi_{\vec{j}} \leqq 1_{\mathcal{O}\backslash\vec{j}}$ so that

$$\begin{aligned}\int_{\mathcal{O}\backslash\vec{j}} \mu &\geqq \int \mu\varphi_{\vec{j}} \\ &= \int \mu\varphi \\ &\geqq \int_{\mathcal{O}} \mu - \delta\end{aligned}$$

with δ being arbitrary, we get the half part of (3.4) for $\mathcal{O}$. By changing $\mathcal{O}$ in $\mathcal{O}+\vec{j}$ and $\vec{j}$ in $-\vec{j}$, we obtain the desired result. Now let K be a compact subset of Y, a positive number, and a function φ in $\mathscr{C}_0(Y)$, $\varphi = 1$ on K such that

$$\int \mu\varphi \leqq \int_{K^\mu} + \delta.$$

The function $\varphi_{+\vec{j}}$ is compactly supported in $Y\backslash\vec{j}$ and $\varphi_{\vec{j}} = 1$ on $K\backslash\vec{j}$. So by definition of $\int_{K\backslash\vec{j}} \mu$

$$\int_{K\backslash\vec{j}} \mu \leqq \int \mu\varphi_{\vec{j}} = \int \mu\varphi \leqq \int_K \mu + \delta.$$

With δ being arbitrary, we get

$$\int_{K\setminus\vec{j}} \mu \leqq \int_K \mu$$

and by a now classical argument

$$\int_K \mu = \int_{K\setminus\vec{j}} \mu.$$

Finally, let us consider a Borel subset B in Y. By definition of $\mu(B)$,

$$\mu(B) = \inf_{\mathcal{O}\supset B} \mu(\mathcal{O}).$$

So let $\mathcal{O}$ be an open subset in Y such that $\mathcal{O}\supset B$ and $\mu(\mathcal{O})\leqq\mu(B)+\delta$. The open set $\mathcal{O}\setminus\vec{j}$ contains $B\setminus\vec{j}$; thus

$$\begin{aligned}\mu(B\setminus\vec{j}) &\leqq \mu(\mathcal{O}\setminus\vec{j})\\ &= \mu(\mathcal{O})\\ &\leqq \mu(B)+\delta.\end{aligned}$$

With δ being arbitrary and by the usual process, we obtain

$$\mu(B) = \mu(B\setminus\vec{j}).$$

Let us now deal with the general case of a Borel set in $\mathbb{R}^N$. It is easy to find a partition $(B_i)_{i\in\mathbb{N}}$ of B in Borel subsets in $\mathbb{R}^N$, and $\vec{j}_i$ in $\mathbb{Z}^N$ such that $B_i\setminus\vec{j}_i\subset Y$. By the additivity of μ on disjoint measurable sets and by the result above for measurable subsets in Y, we get

$$\begin{aligned}\mu(B) &= \sum_i \mu(B_i)\\ &= \sum \mu(B_i-\vec{j}_i)\\ &= \sum \mu((B_i-\vec{j}_i)+(\vec{j}+\vec{j}_i))\\ &= \sum \mu(B_i+\vec{j})\\ &= \mu(B+\vec{j}).\end{aligned} \tag{3.3}$$

PROPOSITION 3.2. *Assume that* $\mu\in M^1_{\text{per}}(Y)$, $\vec{y}\in\,]0,1[^N$ *is such that* $\int_{\partial(y+Y)}|\mu|=0$, *and define* $]0,1\}_i$ *as*

$$]0,1\}_i = \begin{cases}]0,1] & \text{if } y_i>0,\\ ]0,1[& \text{if } y_i=0.\end{cases} \tag{3.4}$$

Then

$$\int_{\vec{y}_0+Y}\mu = \int_{\prod_i]0,1\}_i}\mu. \tag{3.5}$$

Proof. We prove (3.5) by induction on N. For $N=1$ we have to show that for $\alpha\in\,]0,1]$ and for μ in $M^1_{\text{per}}(]0,1[)$, such that $\int_{\{\alpha\}}|\mu|=0$,

$$\int_{]\alpha,\alpha+1[}\mu = \int_{]0,1]}\mu.$$

Toward that aim we write, using Proposition 3.2(ii),

$$\int_{]\alpha,\alpha+1[} \mu = \int_{]0,1]} \mu + \int_{]1,\alpha+1[} \mu$$
$$= \int_{]0,1]} \mu + \int_{]0,\alpha[} \mu$$

and, since μ has no mass on α,

$$\int_{]0,\alpha+1[} \mu = \int_{]0,\alpha[} \mu + \int_{]\alpha,\alpha+1[} \mu.$$

So we get (3.5) for $N=1$. Now let us suppose that (3.5) has been proved for all ν in $M^1_{\text{per}}(]0,1[^{N-1})$, and let μ be in $M^1_{\text{per}}(]0,1[^N)$, and $\vec{y}$ in $]0,1[^N$. We define $\mu_N \in M^1_{\text{per}}(]0,1[)$ by

$$\mu_N(\cdot) = \int_{\vec{y}_N+]0,1[^{N-1}} \mu(x_1, \cdots, x_{N+1})$$

(where $\vec{y}_N = \vec{y} - Y_N \vec{e}_N$). By applying (3.5) to μ_N for $N=1$ we obtain

$$\int_{\vec{y}_N+]0,1[^N} \mu = \int_{\vec{y}_N+]0,1[} \mu_N = \int_{]0,1]} \mu_N$$
$$= \int_{\prod]y_i,y_{i+1}[\times]0,1\}_N} \mu(x_1, \cdots, x_N)$$
$$= \int_{\prod_{1\le i\le N-1}]y_i,y_{i+1}[} \left[\int_{]0,1\}_N} \mu(x_1, \cdots, x_N) \right].$$

Let us now define the measure ν in $M^1_{\text{per}}(]0,1[^{N-1})$ by the formula

$$\langle \nu, \varphi \rangle = \int \mu \varphi(x_1, \cdots, x_{N-1}) \otimes 1_{]0,1[}$$

for every φ in $\mathscr{C}_0(]0,1[^N)$. By applying the induction hypothesis to ν we get

$$\int_{\prod_{1\le i\le N-1}]y_i,y_{i+1}[} \nu = \int_{\prod_{1\le i\le N-1}]0,1]} \nu$$
$$= \int_{\prod_{1\le i\le N-1}]0,1\}_i} \int_{]0,1\}_N} \nu$$
$$= \int_{\prod_{1\le i\le N-1}]0,1\}_i} \mu,$$

which ends the proof of Proposition 3.2.

Remark 3.4. Assume that Ψ is a continuous function defined on X, which is at most linear at infinity, i.e., $|\Psi(\xi)| \le k_1(1+|\xi|)$, for all $\xi \in X$, and such that $\Psi_\infty(\xi) = \lim_{t\to+\infty} (\Psi(t\xi)/t)$ exists in X. Then $\Psi(\mu)$ may be defined as in [14] and [15] by the formula

$$\Psi(\mu) = \Psi \circ g\, dx + \Psi_\infty(\mu^s),$$

where $\mu = g\, dx + \mu^s$ is the Lebesgue decomposition of μ, μ^s being singular. When μ is Y-periodic it is obvious that g and μ^s are Y-periodic, so $\Psi(\mu)$ is too. The same result still holds if we assume that Ψ is convex but not necessarily at most linear at infinity (see [12]) and may be obtained by the use of a duality formula [12]. We now give a result of the approximation in $M^1(Y)$, for a topology related to Ψ.

PROPOSITION 3.3. *Let Ψ be a convex function greater than or equal to zero that is at most linear at infinity, and let μ be in $M^1_{\text{per}}(Y)$. Then there exists u_n in $\mathscr{C}^\infty(\mathbb{R}^N)$, Y periodic, such that*

$$|u_n| \to |\mu| \quad \text{vaguely on } \mathbb{R}^N,$$

$$\Psi(u_n) \to \Psi(\mu) \quad \text{vaguely on } \mathbb{R}^N,$$

$$\int_Y |u_n| \to \int_{Y_1} |\mu|, \qquad \int_Y \Psi(u_n) \to \int_{Y_1} \Psi(\mu)$$

where $Y_1 =]0, 1[^N$.

Proof. Let ρ be in $\mathscr{C}_0^\infty(]-1, +1[^N)\rho \geqq 0, \int \rho = 1, \rho(0) = 1, m \in \mathbb{N}, \rho_m(x) = m^N \rho(mx)$, and $u_m = \rho_m * \mu$. It is easy to see by the definition of $M^1_{\text{per}}(Y)$ that $\rho_m * \mu$ is Y-periodic. Moreover, $|\rho_m * \mu|$ is vaguely convergent toward μ, and by Lemma 2.3 in [11], $\Psi(u_m)$ is vaguely convergent to $\Psi(\mu)$ on $\mathbb{R}^N$. Now let y_0 be in $]0, 1[^N$, such that $\int_{\partial(y_0+Y)} |\mu| = 0$; we have

$$\int_Y \Psi(u_m) = \int_{y_0+Y} \Psi(u_m) \to \int_{y_0+Y} \psi(\mu) = \int_{Y_1} \Psi(\mu),$$

$$\int_Y |u_m| = \int_{y_0+Y} |u_m| \to \int_{y_0+Y} |\mu| = \int_{Y_1} |\mu|,$$

which ends the proof of Proposition 3.3.

3.2. An explicit expression of Ψ_{hom}. The convex set K_{hom}. In this section we give an explicit calculation method of the function Ψ_{hom} for a particular subsequence $\sigma(\varepsilon)$, namely, the sequence $1/n$.

Let us now give several definitions:

$$LU_{\text{per}}(Y) = \{u \in LU_{\text{loc}}(\mathbb{R}^N),\ u(x + \vec{j}) = u(x),$$
$$\text{for almost every } x \text{ in } \mathbb{R}^N \text{ and } \vec{j} \text{ in } \mathbb{Z}^N\};$$

$$BD_{\text{per}}(Y) = \{u \in BD_{\text{loc}}(\mathbb{R}^N),\ u(x + \vec{j}) = u(x)$$
$$\text{for almost every } x \text{ in } \mathbb{R}^N \text{ and } \vec{j} \text{ in } \mathbb{Z}^N\};$$

$$U_{\text{per}}(Y) = \{u \in BD_{\text{per}}(Y),\ \text{div } u \in L^2(\mathbb{R}^N)\}.$$

The internal and external traces for functions of BD, LU, U have been defined by Suquet [17] and Temam [21]. We remark that functions of $LU_{\text{per}}(Y)$ must not have discontinuities on ∂Y. On the contrary, functions of $BD_{\text{per}}(Y)$ are allowed to have one, although the Y-periodicity implies that the external trace $\gamma_+ u$ on some face of Y must be equal to the internal trace $\gamma_- u$ on the opposite face (functions in $U_{\text{per}}(Y)$ must verify, in addition, $u \cdot n^+ = u \cdot n^- / \partial Y$.

By the study made in the first two sections, we know that there exists a subsequence $\sigma(1/n)$ (which we will denote $1/n$ for simplicity) for which the functionals $F_{1/n}(v, \cdot)$ are Γ-convergent to the measure $\Psi_{\text{hom}}(e(u))$, for all $v_n \mapsto u$ in $U(\Omega)$. To show the expression of Ψ_{hom} given in the introduction of § 3, we will need the three lemmas below. Note that the expression found for $\Psi_{\text{hom}}(\xi)$ does not depend on σ. We then may say that the entire sequence $F_{1/n}(\cdot, \varphi)$ is Γ-convergent toward $\Psi_{\text{hom}}(e(u))$.

LEMMA 3.1. *Assume that $\xi \in E$, and define the problem*

$$\inf P_{\text{hom}}(\xi) = \inf_{u \in LU_{\text{per}}(Y)} \left\{ \frac{1}{|Y|} \int_{Y_1} \Psi_{\text{hom}}(\xi + e(u)) \right\}.$$

Then

$$\inf P_{\text{hom}}(\xi) = \Psi_{\text{hom}}(\xi).$$

LEMMA 3.2. *Let $P_n(\xi)$ be the variational problem defined as*

$$\inf P_n(\xi) = \inf_{u \in LU_{\text{per}}(Y)} \left\{ \frac{1}{|Y|} \int_Y \Psi(nx, \xi + e(u)) \right\}.$$

Then

$$\inf P_n(\xi) = \inf P_1(\xi) \quad \forall n \in \mathbb{N}.$$

LEMMA 3.3.

$$\underline{\lim}_{n \to +\infty} \inf P_n(\xi) = \overline{\lim}_{n \to +\infty} \inf P_n(\xi)$$
$$= \inf P_{\text{hom}}.$$

Equation (3.1) follows obviously from these three lemmas. It remains to show them.

Proof of Lemma 3.1. (i) Let ξ be in E and u in $LU_{\text{per}}(Y)$. The equality $\int_Y e(u) = 0$, and the convexity of Ψ_{hom} imply that

$$\Psi_{\text{hom}}(\xi) = \Psi_{\text{hom}} \left\{ \int_Y (\xi + e(u)) \right\}$$
$$\leqq \int_Y \Psi_{\text{hom}}(\xi + e(u))$$
$$= \int_Y \Psi_{\text{hom}}(\xi + e(u)).$$

The reverse inequality is obtained by setting $u = 0$ in the infimum.

(ii) We may suppose without loss of generality that $\xi = 0$. Let u be in $LU_{\text{per}}(Y)$, and u_m as in the proof of Proposition 3.3, i.e., $u_m \in LU_{\text{per}}(Y) \cap \mathscr{C}^\infty(\mathbb{R}^N)$, u_m tends to u in $L^1_{\text{loc}}(\mathbb{R}^N)$, and $|e(u_m)| \to |e(u)|$ vaguely on $\mathbb{R}^N$. By Lemma (2.2) of [14], $\Psi_{\text{hom}}(e(u_m))$ is vaguely convergent toward $\Psi_{\text{hom}}(e(u))$ on $\mathbb{R}^N$. By taking $\vec{y}_0$ such that $\int_{\partial(y_0+Y)} |e(u)| = 0$, we obtain that

$$\int_{Y_1} \Psi_{\text{hom}}(e(u)) = \int_{(y_0+Y)} \Psi_{\text{hom}}(e(u))$$
$$= \lim_{m \to +\infty} \int_{(y_0+Y)} \Psi_{\text{hom}}(e(u_m))$$
$$= \lim_{m \to +\infty} \int_Y \Psi_{\text{hom}}(e(u_m)).$$

Proof of Lemma 3.2. A first step consists of showing that $\inf P_1(\xi) = \inf P_n(\xi)$ for every $n \in \mathbb{N}$. With $\delta > 0$ being given and u_n in $LU_{\text{per}}(Y)$ such that

$$\int_Y \Psi(nx, e(u_n(x) + \xi)\, dx \leqq \inf P_n(\xi) + \delta,$$

we define the function u by the formula

$$u(x) = \frac{1}{n^2} \sum_{1 \leqq j \leqq N} \sum_{0 \leqq i \leqq n-1} u_n\left(\frac{1}{n}(x + i_j)\right).$$

u belongs to $U_{\text{per}}(Y)$, and

$$e(u)(x)=\frac{1}{n^3}\sum_j\sum_i e(u_n)\left(\frac{1}{n}(x+i_j)\right).$$

Since u is admissible for P_1, using the convexity of Ψ, its Y-periodicity, and $Y=\bigcup_j (1/n)(Y+i_j)$, $i_j \leqq n-1$, valid for all $n\in\mathbb{N}$, we get

$$\begin{aligned}
\inf P_1(\xi) &\leqq \int_Y \Psi(x, e(u)(x)+\xi)\\
&\leqq \sum \frac{1}{n^3}\int_Y \Psi\left(x, e(u_n)\left(\frac{1}{n}(x+\vec{i})\right)+\xi\right) dx\\
&= \sum \frac{1}{n^3}\int_{(Y+\vec{i})/n} \Psi[n(x-\vec{i}), e(u_n)(x)+\xi]\, dx\\
&= \sum \int_{(Y+\vec{i})/n} \Psi(nx, e(u_n)(x))\, dx\\
&= \int_Y \Psi(nx, e(u_n)(x)+\xi)\, dx\\
&\leqq \inf P_n(\xi)+\delta.
\end{aligned}$$

For the reverse inequality, let u be in $LU_{\text{per}}(Y)$ so that $\int_Y \Psi(x, e(u)(x)+\xi)\, dx \leqq \inf P_1(\xi)$, and define $u_n(x)=(1/n)u(nx)$. u_n is in $LU_{\text{loc}}(\mathbb{R}^N)$, $(1/n)Y$-periodic, and then Y is admissible as a period, so it is admissible for P_n and we may write

$$\begin{aligned}
\inf P_n(\xi) &\leqq \int_Y \Psi(nx, e(u_n)(x)+\xi)\, dx\\
&= \int_Y \Psi(nx, e(u)(nx)+\xi)\, dx\\
&= \frac{1}{n^3}\int_{nY} \Psi(x, e(u)(x)+\xi)\, dx\\
&= \int_Y \Psi(x, e(u)(x)+\xi)\, dx\\
&\leqq \inf P_1(\xi)+\delta.
\end{aligned}$$

Lemma 3.2 is then proved.

Proof of Lemma 3.3. We first show that $\inf P_{\text{hom}}(\xi) \geqq \overline{\lim}\, \inf P_n(\xi)$. Let u be in $LU_{\text{per}}(Y)$, such that $\int_Y \Psi_{\text{hom}}(e(u)+\xi) \leqq \inf P_{\text{hom}}(\xi)+\delta$. By Proposition 1.2 in § 1, there exists $v_n \in LU(Y)$, $v_n \mapsto u+\xi\cdot x$ in $U(\Omega)$, $\int_Y \Psi(nx, e(v_n)(x)) \mapsto \int_Y \Psi_{\text{hom}}(e(u)+\xi)$, $v_{n|\partial\gamma} = u+\xi\cdot x_{|\partial\gamma}$. Then $u_n = v_n - \xi\cdot x_{|\partial\gamma} = u_{|\partial\gamma}$ and may be extended by periodicity to $\mathbb{R}^N$ as a function of $LU_{\text{per}}(Y)$. We then have

$$\begin{aligned}
\overline{\lim}\, \inf P_n(\xi) &\leqq \overline{\lim} \int_Y \Psi(nx, e(u_n)(x)+\xi)\\
&= \int_Y \Psi_{\text{hom}}(e(u)(x)+\xi)\\
&\leqq \inf P_{\text{hom}}(\xi)+\delta.
\end{aligned}$$

For the reverse inequality, $\delta > 0$ being given, let u_n be in $LU_{\text{per}}(Y)$ such that

$$\frac{1}{|Y|}\int_Y \Psi(nx, e(u_n)(x)+\xi) \leqq \inf P_n + \delta.$$

This implies that $e(u_n)$ is bounded in L^1 and that div u_n is bounded in $L^2(\Omega)$. With u_n being periodic, there exists a constant c_n such that $(u_n - c_n)$ is bounded in $LU_{\text{loc}}(\mathbb{R}^N)$. We may extract from it a subsequence, still denoted $u_n - c_n$, such that

$$(u_n - c_n) \to u \quad \text{in } U_{\text{loc}}(\mathbb{R}^N) \text{ weakly},$$

u being, of course, periodic.

Since $e(u)$ is a bounded periodic measure, there exists $\vec{x}_0, x_0^i \neq 0$, for all i, such that $\int_{\partial\{\vec{x}_0+Y\}} |e(u)| = 0$, which means that u does not present a discontinuity across the boundary of $\vec{x}_0 + Y$. Then, using the properties of Y-periodic measures, we get

$$\begin{aligned}\int_{Y_1} \Psi_{\text{hom}}(e(u)+\xi) &= \int_{\vec{x}_0+Y} \Psi_{\text{hom}}(e(u)+\xi) \\ &\leqq \underline{\lim} \int_{x_0+Y} \Psi(nx, e(u_n)(x)+\xi)\, dx \\ &= \underline{\lim} \int_Y \Psi(nx, e(u_n)(x)+\xi)\, dx,\end{aligned}$$

and then

$$\Psi_{\text{hom}}(\xi) = \inf P_{\text{hom}} \leqq \underline{\lim} \inf P_n.$$

The explicit expression of $\Psi_{\text{hom}}(\xi)$ permits us, first, to find again the explicit expression of Ψ^*_{hom}, and second, to show the equality

$$(\Psi_{\text{hom}})_\infty = (\Psi_\infty)_{\text{hom}} \tag{3.6}$$

where Ψ_∞ denotes the asymptotic function of Ψ. An explicit calculation of Ψ^*_{hom} has been given in [2], [19]; we calculate it by starting from the previous form of Ψ_{hom} and using convex analysis techniques (see Ekeland and Temam [13] for the details of the calculations):

$$\begin{aligned}\Psi^*_{\text{hom}}(\eta) &= \operatorname*{Sup}_\xi \left\{ \xi\cdot\eta - \inf_{u\in LU_{\text{per}}(Y)} \left\{ \int_Y \Psi(x, e(u)+\xi) \right\} \right\} \\ &= -\inf_{\xi, u\in LU_{\text{per}}(Y)} \left\{ \int_Y \Psi(x, e(u)+\xi) - \xi\cdot\eta \right\} \\ &= -\operatorname*{Sup}_{\substack{\text{div } \sigma = 0 \\ \int_Y \sigma = 0 \\ \sigma \in L^2_{\text{per}}(Y) \\ (\sigma+\eta)^D \in K^D(x)}} \left\{ -\int_Y \Psi^*(x, \sigma+\eta) \right\}\end{aligned}$$

$$\Psi^*_{\text{hom}}(\eta) = \inf_{\substack{\text{div } \alpha = 0 \\ \int_Y \sigma = 0}} \left\{ \int_Y \Psi^*(x, \sigma+\eta) \right\}. \tag{3.7}$$

This also gives a rather precise description of the convex set K_{hom}:

$$K_{\text{hom}} = \text{Dom } \Psi^*_{\text{hom}} = \left\{ \eta, \exists \sigma \in L^2_{\text{per}}(Y), (\sigma+\eta)^D \in K^D(x), \int_Y \sigma = 0, \text{div } \sigma = 0 \right\} \tag{3.8}$$

where $K^D(x) = K(x) \cap \{\sigma, \text{tr } \sigma = 0\}$.

3.3. The functions $(\Psi_{\text{hom}})_\infty$ and $(\Psi_\infty)_{\text{hom}}$. In this section, the results of §§ 1 and 2 concerning Ψ are extended to functions that do not satisfy (1.1) and (1.2); moreover, the equality $(\Psi_{\text{hom}})_\infty = (\Psi_\infty)_{\text{hom}}$ is shown. It permits us to prove the approximation result (3.11), which is not obvious.

Let us recall that when Ψ is convex and continuous, its asymptotic function Ψ_∞ is defined as

$$\Psi_\infty(\xi) = \lim_{t \to +\infty} \frac{\Psi(t\xi)}{t}.$$

If Ψ satisfies (1.2), the domain of Ψ_∞ is the space of deviators $E^D = \{\xi \in E, \operatorname{tr} \xi = 0\}$ and Ψ_∞ verifies $\Psi_\infty(x, 0) = 0$, and when ξ belongs to E^D,

$$C_0|\xi^D| \leqq \Psi_\infty(x, \xi) \leqq C_1|\xi^D|.$$

We define the spaces

$$LU^0(\Omega) = \{u \in LU(\Omega), \operatorname{div} u = 0\}, \qquad U^0(\Omega) = \{u \in U(\Omega), \operatorname{div} u = 0\}$$

and the analogues of $\bar{\Psi}^-(u)$, $\bar{\Psi}^+(u)$:

$$\langle \bar{\Psi}^-_\infty(u)\varphi \rangle = \Gamma^-(L^p) \varliminf_{\substack{\varepsilon \to 0 \\ v \to u}} F_{\infty\varepsilon}(u, \varphi), \tag{3.9}$$

$$\langle \bar{\Psi}^+_\infty(u), \varphi \rangle = \Gamma^-(L^p) \varlimsup_{\varepsilon \to 0} F_{\infty\varepsilon}(u, \varphi) \tag{3.10}$$

where $\varphi \geqq 0$ is in $\mathscr{C}(\bar{\Omega})$ and

$$F_{\infty\varepsilon}(v, \varphi) = \begin{cases} \displaystyle\int \Psi_\infty\left(\frac{x}{\varepsilon}, e(v_\varepsilon)(x)\right) \varphi(x)\, dx & \text{if } u_\varepsilon \in LU^0(\Omega), \\ +\infty & \text{if not,} \end{cases}$$

and we have the analogue of Theorem 1.1.

THEOREM. (i) *There exists a subsequence ε' for which $\bar{\Psi}^-_\infty(u) = \bar{\Psi}^+_\infty(u) = \bar{\Psi}_\infty(u)$ for every u in $U^0(\Omega)$; $\bar{\Psi}_\infty(u)$ may be extended as a bounded measure on Ω, absolutely continuous with respect to $|e^D(u)|$.*

(ii) *$\bar{\Psi}_\infty$ is a convex function of u and verifies*

$$|\bar{\Psi}_\infty(u) - \bar{\Psi}_\infty(v)| \leqq C_1|e^D(u) - e^D(v)|.$$

Proof. The proof follows that of Theorem 1.1, and the changes it brings are obvious if we use at each step the approximant $\bar{u}_\varepsilon$, the following result of approximation, also stated in [21]. Let Ω be an open C^1 bounded set of $\mathbb{R}^N$; for every u in $U^0(\Omega)$ there exists $\bar{u}_\varepsilon$ in $C^1(\Omega, E) \cap L^1(\Omega, E)$ such that

$$\bar{u}_\varepsilon \to u \quad \text{in } L^1(\Omega),$$

$$|e^D(\bar{u}_\varepsilon)| \to |e^D(u)| \quad \text{tightly on } \Omega,$$

$$\operatorname{div} \bar{u}_\varepsilon = 0,$$

$$\bar{u}_\varepsilon = u_{|\Gamma}.$$

This is the main argument used to prove the approximation result below. It is left to the reader:

(3.11) Let u be in $LU^0(\Omega)$. There exists u_ε in $LU^0(\Omega)$, u_ε tends to u in $U(\Omega)$, $u_\varepsilon = u_{|\Gamma}$, and $\int_\Omega \Psi_\infty(x/\varepsilon, e(u_\varepsilon)) \to \int_\Omega \bar{\Psi}_\infty(u)$.

Following the argument in Proposition 2.1, it is easy to show that there exists a punctual function that we denote $(\Psi_\infty)_{\text{hom}}$, of domain E^D such that

$$\langle \bar{\Psi}_\infty(u), \varphi \rangle = \int (\Psi_\infty)_{\text{hom}}(e(u))(x)\varphi(x)\, dx$$

for every u in $LU^0(\Omega)$, φ in $\mathscr{C}_0(\Omega)$; of course, the results in Theorem 2.1 and Proposition 2.3 are still valid for $\Psi_\infty(u)$ when u belongs to $LU^0(\Omega)$.

We may give an explicit form of $(\Psi_\infty)_{\text{hom}}$. For $\xi \in E^D$ it is

$$(\Psi_\infty)_{\text{hom}}(\xi) = \inf_{u \in LU_{\text{per}}(Y)} \left(\int_Y \Psi_\infty(x, \xi + e(u)) \right). \tag{3.12}$$

Note that the domain of the function in the right member of (3.12) is E^D. Indeed, the infimum must be taken for u verifying $\operatorname{tr} \xi + \operatorname{div} u = 0$. But $u \in LU_{\text{per}}(Y)$ implies that $\int_Y \operatorname{div} u = 0$, and then $\operatorname{tr} \xi = \int_Y \operatorname{tr} \xi = \int_Y \operatorname{tr} \xi + \operatorname{div} u = 0$.

To show (3.12), we proceed as we did for Ψ (the analogues of Lemmas 3.1–3.3 are true and may be proved by adapting the arguments used already).

Now we will prove the following:

$$(\Psi_{\text{hom}})_\infty = (\Psi_\infty)_{\text{hom}}. \tag{3.13}$$

The left member of (3.13) makes sense because it is the asymptotic function of Ψ_{hom}. Moreover, (1.2) yields that $\Psi_{\text{hom}}(\xi)$ is "quadratic" with respect to $\operatorname{tr} \xi$, and "linear" with respect to $|\xi^D|$, and then $\operatorname{dom} (\Psi_{\text{hom}})_\infty = E^D$. Since $(\Psi_{\text{hom}})_\infty$ and $(\Psi_\infty)_{\text{hom}}$ are lower semicontinuous and proper, it suffices to show that their conjugates coincide. We have

$$\begin{aligned}(\Psi_{\text{hom}})^*_\infty(\eta) &= \sup_{\xi \in E^D} \left\{ \eta \cdot \xi - \inf_{\substack{u \in LU_{\text{per}}(Y) \\ \operatorname{div} u = 0}} \left(\int_Y \Psi_\infty(x, \xi + e(u)) \right) \right\} \\ &= - \inf_{\substack{\xi \in E^D \\ u \in LU^0_{\text{per}}(Y)}} \left\{ \int_Y \Psi_\infty(x, \xi + e(u)) - \eta \cdot \xi \right\}.\end{aligned}$$

These calculations are rather technical, although classical in the theory of duality (cf. Ekeland and Temam [13]). We obtain

$$\Psi^*_{\text{hom}}(\eta) = \sup_{\substack{\exists \sigma \in L^2_{\text{per}}(Y) \\ \operatorname{div} \sigma = 0 \text{ in } \mathbb{R}^N \\ \int_Y \sigma^D = 0 \\ (\sigma + \eta)^D \in K^D}} \{0\} = \chi_{\overline{K_{\text{hom}}}}(\eta)$$

where $\overline{K_{\text{hom}}} = \{\eta, \exists \sigma \in L^2_{\text{per}}(\chi), \operatorname{div} \sigma = 0 \text{ in } \mathbb{R}^N, \int_Y \sigma^D = 0, (\sigma(x) + \eta)^D \in K^D(x)\}$.

Let us note that $\overline{K_{\text{hom}}} = K_{\text{hom}}$. First, we obviously have $K_{\text{hom}} \subset \overline{K_{\text{hom}}}$. Second, take η in $\overline{K_{\text{hom}}}$, and $\sigma \in L^2_{\text{per}}(Y)$, such that $\operatorname{div} \sigma = 0$ in $\mathbb{R}^N$, $\int_Y \sigma^D = 0$, $(\sigma(x) + \eta)^D \in K^D(x)$. Then $\sigma_{1ij} = (\sigma_{ij} - \int_Y \operatorname{tr} \sigma \delta_{ij})$ verifies $\sigma_1 \in L^2_{\text{per}}(Y)$, $\operatorname{div} \sigma_1 = 0$, $\int_Y \sigma_1 = 0$, $(\sigma + \eta)^D \in K^D(x)$, and thus $\eta \in K_{\text{hom}}$. We have finally obtained that

$$(\Psi_{\text{hom}})_\infty = (\Psi_\infty)_{\text{hom}}.$$

In particular, for u in $U^0(\Omega)$, there exists $u_\varepsilon \in LU(\Omega)$, $\operatorname{div} u_\varepsilon = 0$, $u_\varepsilon \to u$ in $L^1(\Omega)$, $u_\varepsilon = u_{|\Gamma}$ such that

$$\int_\Omega \left(\Psi_\infty\left(\frac{x}{\varepsilon}, e(u_\varepsilon) \right)(x)\, dx \right) \to \int_\Omega (\Psi_{\text{hom}})_\infty(e(u)).$$

This result will be important to study the relations between the prehomogenized limit analysis problems and the homogenized one in [9]. □

In [9] we apply these mathematical results to homogenization in plasticity. The displacement problem for a nonhomogeneous, elastic, perfect plastic material made with a great number of identical cells of size ε may be written in the following variational form:

$$\inf_{\substack{u\in H^1\\ u=u_0|\Gamma_0}} \left\{ \int_\Omega \psi\left(\frac{x}{\varepsilon}, e(u)\right) - \lambda \int_\Omega fu - \lambda \int_{\Gamma_1} Fu \right\}$$

where $u_0 \in H^{1/2}(\Gamma)$, $u_0 = \chi_{\Gamma_0} u_0$, f is the body force ($f \in L^{N/(N-1)}$), F is the boundary force ($F \in L^\infty(\Gamma_1)$), and λ is a real parameter; Ψ is the energy functional that verifies the assumptions in § 1. In [9] we show that when λ is less than a positive limit load λ_0, which we define precisely, then inf P_ε converges toward the homogenized problem

$$\inf_{\substack{u\in H^1(\Omega)\\ u=u_0|\Gamma_0}} \left\{ \int_\Omega \psi_{\text{hom}}(e(u)) - \lambda \int_\Omega fu - \lambda \int_{\Gamma_1} Fu \right\}$$

where ψ_{hom} is the function defined in § 1.

Appendix. Let us give a concrete mechanical example for which the formulas found in § 3 give an explicit description of the homogenized convex set, and of Ψ_{hom}. Suppose that k_{ij} is a divergence-free function, with $k_{ii} = 0$, $k_{12}(x_1, x_2, x_3) = k_1(x_3)$, $k_{13}(x_1, x_2, x_3) = k_2(x_2)$, $k_{23}(x_1, x_2, x_3) = k_3(x_1)$, k_1, k_2, k_3 being periodic of period 1. We assume in addition that

$$0 < \alpha_0 \leqq \sum_j |k_{ij}(x)| \leqq \alpha_1 < +\infty.$$

We have div $(k) = 0$ and $K(x)$ is defined as

$$K(x) = \{\sigma, |\sigma_{ij}^D|(x) \leqq k_{ij}(x)\}.$$

By the formula following (3.7) in § 3, K_{hom} is defined as $K_{\text{hom}} = \{\eta \in E, \exists \sigma \in L^2_{\text{per}}(Y), \int_Y \sigma = 0, \text{ div } \sigma = 0, \text{ and } \sigma + \eta \in K\}$. We will show that

$$K_{\text{hom}} = \left\{ \eta \in E, |\eta_{ij}^D| \leqq \int_Y (k_{ij})(x) \right\}.$$

Toward that aim, we denote by $\overline{K_{\text{hom}}}$ the right member in the equality above; we begin by showing that $K_{\text{hom}} \subset \overline{K_{\text{hom}}}$. Thus suppose that $\eta \in K_{\text{hom}}$, and that σ verifies $\sigma \in L^2_{\text{per}}(Y)$, $\int_Y \sigma = 0$, $\sigma + \eta \in K$, and div $\sigma = 0$. We have

$$|\eta_{ij}^D| = \left| \int_Y (\sigma + \eta)_{ij}^D \right| \leqq \int_Y |(\sigma + \eta)_{ij}^D| \leqq \int_Y k_{ij}(x)$$

and then $\eta \in \overline{K_{\text{hom}}}$. Conversely, if η is such that $|\eta_{ij}^D| \leqq \int_Y k_{ij}(x)$, let α_{ij} be $\eta_{ij}^D / \int_Y k_{ij}(x)$; we have $|\alpha_{ij}| \leqq 1$ for all ij, and we define $\sigma_{ij} = \alpha_{ij} k_{ij} - \eta_{ij}^D$. We have $\int_Y \sigma_{ij} = 0$ $(\sigma + \eta)_{ij}^D = (\sigma_{ij} + \eta_{ij}^D)(x) = \alpha_{ij} k_{ij}(x) \leqq k_{ij}(x)$, and div $\sigma = 0$; this implies that $\eta \in K_{\text{hom}}$.

Suppose now that

$$\text{(A1)} \qquad \Psi^*(x, \xi) = \begin{cases} \frac{1}{2}\alpha (\text{tr } \xi)^2 & \text{if } \xi^D \in K^D(x), \\ +\infty & \text{if not} \end{cases}$$

where α is some positive constant. Then $\Psi(x, \cdot)$ satisfies the hypothesis in § 1. Let us calculate Ψ^*_{hom}. Using (3.7) we have

$$\text{(A2)} \qquad \Psi^*_{\text{hom}}(\xi) = \inf_{\substack{\text{div } \sigma = 0\\ \int_Y \sigma = 0\\ (\xi+\sigma)^D \in K^D(x)}} \left(\frac{1}{2} \alpha \int_Y (\text{tr } (\xi + \sigma))^2 \right).$$

Let us show that

$$\Psi^*_{\text{hom}}(\xi)=\begin{cases}\frac{1}{2}\alpha(\operatorname{tr}\xi)^2 & \text{if } |\xi^D_{ij}|\leqq\int_Y k_{ij}(x),\\ +\infty & \text{if not.}\end{cases}$$

Since the two sides of (A2) are infinite when $\xi\notin K_{\text{hom}}$, we may suppose that $\xi\in K_{\text{hom}}$. By setting $\sigma=0$ in the right-hand side of (A2) we get that

$$\Psi^*_{\text{hom}}(\xi)\leqq\tfrac{1}{2}\alpha(\operatorname{tr}\xi)^2.$$

Now Jensen's inequality yields

$$\frac{1}{2}\alpha\left(\int_Y(\operatorname{tr}(\xi+\sigma))\right)^2\leqq\frac{1}{2}\int_Y\alpha(\operatorname{tr}(\xi+\sigma))^2$$

and then, with $\int_Y\operatorname{tr}\sigma=0$,

$$\frac{1}{2}\alpha\int_Y(\operatorname{tr}(\xi))^2\leqq\inf_{\substack{\sigma\\ \int_Y\sigma=0}}\left(\frac{1}{2}\int_Y\operatorname{tr}(\xi+\sigma)^2\right)$$

$$=\Psi^*_{\text{hom}}(\xi).$$

We have finally obtained the desired result, and are able to calculate Ψ_{hom}:

$$\Psi_{\text{hom}}(\xi)=\frac{1}{2\alpha}(\operatorname{tr}\xi)^2+\sum_{ij}\left(\int_Y k_{ij}(x)\right)|\xi^D_{ij}|.$$

Acknowledgments. The authors thank the referee for remarks and advice, and G. Bouchitte, François Murat, and Colette Picard for fruitful discussions on this subject.

REFERENCES

[1] H. Attouch, *Variational Convergence for Functions and Operators*, Pitman, London, 1984.

[2] G. Bouchitte, *Convergence et relaxation de fonctionnelles du calcul des variations à croissance linéaire. Application à l'homogénéisation en plasticité*, Publ. Avamac 85-10, Université de Perpigran, Perpignan, France, 1985.

[3] ———, *Homogénéisation sur $BV(\Omega)$ de fonctionnelles intégrales à croissance linéaire. Application à un problème d'analyse limite en plasticité*, C.R. Acad. Sci. Sér. I Math., 301 (1985), pp. 785-788.

[4] G. Buttazo and G. Dal Maso, *Γ-limit of integral functionals*, J. Analyse Math., 37 (1980), pp. 147-185.

[5] P. L. Carbone and C. Sbordone, *Some properties of Γ-limit of integral functionals*, Ann. Mat. Pura Appl. (4), 2 (1979), pp. 1-60.

[6] G. Dal Maso and L. Modica, *Nonlinear stochastic homogenization*, Ann. Mat. Pura Appl. (4), 144 (1986), pp. 347-389.

[7] E. De Giorgi, *Γ-convergenza e G-convergenza*, Boll. Un. Mat. Ital. A (6), 14 (1977), pp. 213-220.

[8] E. De Giorgi and Franzoni, *Su un tipo di convergenza variazional*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. 8, 58 (1975), pp. 842-850.

[9] F. Demengel and Tang Qi, *Homogenization for elastic perfect plasticity*, Prépublication, Université Paris XI, Orsay, France.

[10] F. Demengel, *Some compacity theorems in spaces of functions with bounded derivatives and application to plasticity*, Arch. Rational Mech. Anal., 105 (1989), pp. 123-161.

[11] F. Demengel and R. Temam, *Convex function of a measure*, Indiana Univ. Math. J., 33 (1984), pp. 673-709.

[12] ———, *Convex function of a measure, the unbounded case*, Actes des journées de Fermat 1985, North-Holland, Amsterdam, New York, 1986.

[13] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, New York, 1976.

[14] M. GIAQUINTA AND G. MODICA, *Nonlinear systems of the type of the stationary Navier-Stokes system*, Ann. Mat. Pura Appl. (4), 149 (1987), pp. 41-59.

[15] R. V. KOHN AND R. TEMAM, *Dual spaces of stresses and displacement with applications to Hencky plasticity*, Appl. Math. Optim., 10 (1983), pp. 1-35.

[16] P. MARCELLINI, *Periodic solutions and homogenization of nonlinear variational problems*, Ann. Mat. Pura Appl. (4), 117 (1978), pp. 139-152.

[17] P. SUQUET, *Plasticité et homogénéisation*, Thèse de doctorat d'état, Université de Paris VI, Paris, 1982.

[18] ———, *Analyse limite et homogénéisation*, C.R. Acad. Sci. Paris Sér. II Math., 296 (1983), pp. 1355-1358.

[19] TANG QI, *Homogénéisation des problèmes élastiques en dimension* 3, Thèse d'Université, Université Paris 11, Paris, 1986.

[20] L. TARTAR, Cours Peccot au Collège de France, 1977, unpublished.

[21] R. TEMAM, *Problèmes mathématiques en plasticité*, Gauthiers-Villars, Paris, 1983.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 436-460, March 1990
© 1990 Society for Industrial and Applied Mathematics
009

SOME NONLINEAR WEAK ERGODIC THEOREMS*

ROGER D. NUSSBAUM†

Abstract. Let C be a cone with nonempty interior in a Banach space and, for $j \geqq 1$, let $f_j : \mathring{C} \to \mathring{C}$ be a sequence of maps. It is frequently assumed that each f_j is homogeneous of degree 1 and order-preserving with respect to the partial ordering induced by C; but it is not assumed that $f_j(C-\{0\}) \subset \mathring{C}$. If $F_m = f_m f_{m-1} \cdots f_1$, the composition of the first m f_j, and if d denotes Hilbert's projective metric, then theorems (usually called weak ergodic theorems in the population biology literature) can be proved ensuring that, for all x and y in $\mathring{C}$, $\lim_{m\to\infty} d(F_m(x), F_m(y)) = 0$ and (if C is normal) $\lim_{m\to\infty} \|(F_m(x)/\|F_m(x)\|) - (F_m(y)/\|F_m(y)\|)\| = 0$. If $u \in \mathring{C}$ is fixed and assumptions on the f_j are strengthened, it can be proved that for every $z \in \mathring{C}$ there exists $\lambda(x) > 0$ such that $\lim_{m\to\infty} \|F_m(x) - \lambda(x) F_m(u)\| = 0$. These theorems are applied to the case where $C = \{x \in \mathbb{R}^n : x_i \geqq 0 \text{ for } 1 \leqq i \leqq n\}$ and where the maps f_j belong to a class M arising in the theory of "means and their iterations" and in certain problems from population biology.

Key words. Hilbert's projective metric, nonlinear weak ergodic theorems, cones, nonlinear cone maps, positive linear operators

AMS(MOS) subject classifications. 47A35, 47B55, 47H07, 47H09

1. Preliminaries. In an effort to make this paper self-contained, we begin by recalling some definitions and theorems from the literature. By a cone C (with vertex at 0) in a Banach space X we mean a closed, convex subset C of X such that (a) $tC \subset C$ for all $t \geqq 0$ and (b) if $x \in C - \{0\}$, then $-x \notin C$. A cone induces a partial ordering on X by

$$x \leqq y \quad \text{if and only if } y - x \in C.$$

Two elements $x, y \in C$ will be called "comparable" if there exist positive reals α and β such that

$$\alpha x \leqq y \leqq \beta x, \qquad \alpha, \beta > 0.$$

If x and y in C are comparable, we follow Bushell [6] and define

$$M(y/x) = \inf\{\beta > 0 : y \leqq \beta x\}, \tag{1.1}$$

$$m(y/x) = \sup\{\alpha > 0 : \alpha x \leqq y\}. \tag{1.2}$$

If $u \in C - \{0\}$, C_u will denote the set of elements of C that are comparable to u. If u is an element of the interior of C, C_u is the interior of C. In general C_u satisfies all properties of a cone except closedness.

Associated to the set C_u is a natural normed linear space E_u,

$$E_u = \{x \in X : \text{there exists } \alpha > 0 \text{ such that } -\alpha u \leqq x \leqq \alpha u\}.$$

For $x \in E_u$, we define a norm $|x|_u$ by

$$|x|_u = \inf\{\alpha > 0 : -\alpha u \leqq x \leqq \alpha u\}.$$

A cone C in a Banach space X is called "normal" if there exists a constant M such that

$$\|x\| \leqq M\|y\|$$

* Received by the editors July 5, 1988; accepted for publication (in revised form) May 3, 1989. This research was partially supported by National Science Foundation grants DMS 85-03316 and DMS 88-05395.

† Mathematics Department, Rutgers University, New Brunswick, New Jersey 08903.

for all $x, y \in C$ such that $x \leqq y$. A cone C is "total" if the closure of the linear span of C equals X. If C is a cone, C^* will always denote the set of continuous linear functionals $\psi \in X^*$ such that $\psi(x) \geqq 0$ for all $x \in C$. It is not hard to see that if C is a total cone, then C^* is a cone.

The basic technical tool we will use in this paper is the so-called "Hilbert's projective metric" d. If C is a cone in a Banach space X and $u \in C - \{0\}$, then for $x, y \in C_u$, define $\beta = M(y/x)$, $\alpha = m(y/x)$, and

$$d(x, y) = \log\left(\frac{\beta}{\alpha}\right), \tag{1.3}$$

$$\bar{d}(x, y) = \log(\max(\beta, \alpha^{-1})). \tag{1.4}$$

We can easily prove (see [6]) that the projective metric d satisfies all properties of a metric except that $d(y, x) = 0$ if and only if $y = \lambda x$ for some $\lambda > 0$. On the other hand, $\bar{d}$, which was introduced by Thompson [36], is a metric on C_u. If $\Sigma = \{x \in C_u : \|x\| = 1\}$, (Σ, d) is a metric space, and it is natural to ask if (Σ, d) is complete. It is proved in [37] that (Σ, d) is complete if and only if

$$\sup\{\|x\| : 0 \leqq x \leqq u\} < \infty, \tag{1.5}$$

and (1.5) is satisfied if and only if E_u is a complete normed linear space. Also, one can show that $(C_u, \bar{d})$ is complete if and only if (1.5) is satisfied. It should be noted that the results in [37] are closely related to much earlier theorems of Thompson [36] and Birkhoff (see Theorem 5 in [5] and Remark 1.1 in [25]). Notice also that if $\psi \in C^*$ and $\psi(u) > 0$ and

$$\Sigma_\psi = \{x \in C_u : \psi(x) = 1\},$$

then (Σ_ψ, d) is complete if and only if (Σ, d) is complete, because (Σ_ψ, d) and (Σ, d) are isometric.

If $K = \{x \in \mathbb{R}^n : x_i \geqq 0 \text{ for } 1 \leqq i \leqq n\}$, K will be called the "standard cone in $\mathbb{R}^n$." Obviously, K is normal, so if $\Sigma_1 = \{x \in \mathring{K} : \sum_{i=1}^n x_i = 1\}$ or $\Sigma_2 = \{x \in \mathring{K} : x_1 = 1\}$, the remarks above show that (Σ_1, d) and (Σ_2, d) are complete.

We also need to recall some results about positive linear operators. Suppose that C is a cone in a Banach space X and that $L : X \to X$ is a bounded linear operator such that $L(C) \subset C$. Assume that Lx and Ly are comparable for all $x, y \in C$ such that $Lx \neq 0$ and $Ly \neq 0$ and define a number $\Delta(L)$, the "projective diameter of $L(C) - \{0\}$" by

$$\Delta(L) = \sup\{d(Lx, Ly) : x, y \in C, Lx \neq 0 \text{ and } Ly \neq 0\}. \tag{1.6}$$

If $Lx = 0$ for all $x \in C$, we define $\Delta(L) = 0$. If L is as above and $\Delta(L) < \infty$ we shall say that "L has finite projective diameter."

If $x, y \in C - \{0\}$ are not comparable, define $d(x, y) = \infty$. If $x, y \in C - \{0\}$ and $M(y/x) < \infty$, define

$$\operatorname{osc}(y/x) = M(y/x) - m(y/x). \tag{1.7}$$

If $M(y/x) = \infty$, define $\operatorname{osc}(y/x) = \infty$. If L is a bounded linear operator such that $L(C) \subset C$ and Lx and Ly are nonzero and comparable for all $x, y \in C - \{0\}$, define

$$k(L) = \inf\{k > 0 : d(Lx, Ly) \leqq k d(x, y) \text{ for all } x, y \in C - \{0\}\},$$

$$N(L) = \inf\{\lambda > 0 : \operatorname{osc}(Ly/Lx) \leqq \lambda \operatorname{osc}(y/x) \text{ for } x, y \in C - \{0\}\}.$$

It is easy to see that $k(L) \leqq 1$ and $N(L) \leqq 1$. However, if $\Delta(L) < \infty$, results of Birkhoff [4], [5] and Hopf [19], with refinements of Ostrowski [27], Bauer [2], Bushell [6], [7], and others [20], [37] imply that

$$N(L) = k(L) = \tanh\left(\frac{\Delta(L)}{4}\right) < 1. \tag{1.8}$$

As a particular example, note that if K is the standard cone in $\mathbb{R}^n$ and L is an $n \times n$ matrix, all of whose entries are positive, then $\Delta(L) < \infty$. In fact, it is not hard to prove that $\Delta(L) = \sup_{i,j} d(Le_i, Le_j)$, where e_i, $1 \leqq i \leqq n$, is the standard basis of $\mathbb{R}^n$. From this observation and (1.8) we derive an explicit formula (see [6], [35]) for $\Delta(L)$ and $k(L)$.

If C is a cone and $D \subset C$, a map $f: D \to C$ will be called nonexpansive with respect to d if

$$d(f(x), f(y)) \leqq d(x, y) \quad \text{for all } x, y \in D. \tag{1.9}$$

We have the obvious modification for $\bar{d}$. A map $f: D \to C$ will be called "order-preserving" if $f(x) \leqq f(y)$ for all $x, y \in D$ such that $x \leqq y$. The map f will be called "homogeneous of degree 1" on D if

$$f(tx) = tf(x) \quad \text{for all } t > 0 \quad \text{and} \quad x \in D \quad \text{such that } tx \in D,$$

and will be called "subhomogeneous" on D if $f(tx) \geqq tf(x)$ for all t, $0 < t \leqq 1$, and $x \in D$ such that $tx \in D$. It is an easy exercise that if $u \in C - \{0\}$, $D = C_u$ and $f: D \to D$ is order preserving and homogeneous of degree 1, then f is nonexpansive with respect to d: see [6], [25], [29]. Thompson [36] observed that, if $f: C_u \to C_u$ is subhomogeneous and order-preserving, then f is nonexpansive with respect to $\bar{d}$. Potter [29] observed that, for $\psi \in C^*$ with $\psi(u) > 0$, the restriction of f to $\Sigma_\psi = \{x \in C_u: \psi(x) = 1\}$ is nonexpansive with respect to d.

Now suppose that C is a cone in a Banach space X, $u \in C - \{0\}$, and S is a collection of maps $f: C_u \to C_u$. In most of this paper we will assume that f is order preserving and homogeneous of degree 1 for every $f \in S$. Suppose that $f_j \in S$, $1 \leqq j < \infty$, is a sequence of functions in S and define

$$F_n = f_n f_{n-1} f_{n-2} \cdots f_1 \tag{1.10}$$

for $n \geqq 1$. We are interested in finding further conditions ensuring that for all $x, y \in C_u$,

$$\lim_{n \to \infty} d(F_n(x), F_n(y)) = 0. \tag{1.11}$$

Such results are called "weak ergodic theorems" in the population biology literature [11], [17]. The linear theory is well understood: see the excellent survey article [11] by Cohen. If (1.5) is satisfied, it is known (see eq. (1.20a) in [25]) that there exists a constant M such that

$$\begin{aligned} &\|x - y\| \leqq M[\exp(d(x, y)) - 1] \quad \text{for all } x, y \in \Sigma, \\ &\Sigma = \{x \in C_u: \|x\| = 1\}. \end{aligned} \tag{1.12}$$

Using (1.12) we can see that if (1.5) and (1.11) hold and the functions f_j are homogeneous of degree 1, then

$$\lim_{n \to \infty} \|F_n(x)\| F_n(x)\|^{-1} - F_n(y)\| F_n(y)\|^{-1}\| = 0. \tag{1.13}$$

Note that if (1.5) is satisfied, $C \cap E_u$ is a normal cone with nonempty interior C_u in the Banach space E_u, so by working in E_u we can assume that C is normal with nonempty interior.

In fact, the question we are asking is motivated by a particular class of maps $\mathcal{M}$ defined on the interior of the standard cone K in $\mathbb{R}^n$, so we recall the definition of $\mathcal{M}$ (see [24], the Introduction to [25], and § 4 of [23]). Recall that a probability vector σ is a vector $\sigma \in K$ such that $\sum_{i=1}^{n} \sigma_i = 1$. If r is a real number and σ a probability vector, define $M_{r\sigma} : \mathring{K} \to \mathbb{R}$ by

$$M_{r\sigma}(x) = \left(\sum_{i=1} \sigma_i x_i^r \right)^{1/r}.$$

If $r = 0$, define

$$M_{0\sigma}(x) = \prod_{i=1}^{n} x_i^{\sigma_i}.$$

Such maps, of course, have an extensive classical theory, described in [18]. For $1 \leqq i \leqq n$, let Γ_i be a finite collection of ordered pairs (r, σ), with r a real number and σ a probability vector. For $(r, \sigma) \in \Gamma_i$ let $c_{ir\sigma}$ be a positive real number and define $f_i : \mathring{K} \to (0, \infty)$ by

$$f_i(x) = \sum_{(r,\sigma)\in\Gamma_i} c_{ir\sigma} M_{r\sigma}(x). \tag{1.14}$$

Define f_i to be the ith component of a map $f : \mathring{K} \to \mathring{K}$. If $f : \mathring{K} \to \mathring{K}$ can be written in this form, we will write $f \in M$. If f can be written as in (1.14) in such a way that $r \geqq 0$ for all $(r, \sigma) \in \Gamma_i$ and $1 \leqq i \leqq n$, we say that $f \in M_+$; if f can be written as in (1.14) such that $r < 0$ for all $(r, \sigma) \in \Gamma_i$ and $1 \leqq i \leqq n$, we say that $f \in M_-$. Note that if f is a linear map such that $f(\mathring{K}) \subset \mathring{K}$, then $f \in M_+ \cap M_-$. We define $\mathcal{M}(\mathcal{M}_+, \mathcal{M}_-)$ to be the smallest set of maps $f : \mathring{K} \to \mathring{K}$ such that $\mathcal{M} \supset M (\mathcal{M}_+ \supset M_+, \mathcal{M}_- \supset M_-)$ and $\mathcal{M}(\mathcal{M}_+, \mathcal{M}_-)$ is closed under addition of functions, composition of functions, and multiplication by positive scalars. The class $\mathcal{M}$ arises in the theory of "means and their iterations"; see [1], [12], [15], [23]-[26]. It is proved in [26] (this is not hard) that if $f \in \mathcal{M}$, then f extends continuously to K and $f | \mathring{K}$ is C^∞. We will see that establishing weak ergodic theorems may already be nontrivial when $S = \mathcal{M}$ and $C = K$. Existing nonlinear weak ergodic theorems as, for example, in the work of Fujimoto and Krause [16], are frequently inapplicable. On the other hand, we have not attempted to give an all-inclusive abstract framework: there are examples to which our general theorems are not directly applicable but which can be handled with theorems from [16].

2. Some nonlinear weak ergodic theorems. The following definition will play a crucial role in our subsequent work.

DEFINITION 2.1. Let C be a cone in a Banach space X and D a subset of C such that all elements of D are comparable. Suppose that $f_j : D \to D$, $j \geqq 1$, is a sequence of maps and define $F_m = f_m f_{m-1} \cdots f_1$ to be the composition of the first m functions f_j, $1 \leqq j \leqq m$. We say that $\langle f_j \rangle$ has "the bounded orbit property" (with respect to Hilbert's projective metric) if for every $x \in D$, there exist $y \in D$ and $R > 0$ (possibly depending on x) such that

$$d(F_m(x), y) \leqq R \quad \text{for all } m \geqq 1.$$

If each of the functions f_j is nonexpansive with respect to d, it is an easy exercise (left to the reader) to prove that $\langle f_j \rangle$ has the bounded orbit property if and only if there exist $x_0, y_0 \in D$ and $R_0 > 0$ such that

$$d(F_m(x_0), y_0) \leqq R_0 \quad \text{for all } m \geqq 1.$$

If $D = C_u$ for some $u \in C - \{0\}$ and each of the functions $f_j : D \to D$ in Definition 2.1 is homogeneous of degree 1 and order-preserving, select $\psi \in C^*$ such that $\psi(u) > 0$ and define $\Sigma_\psi = \{x \in C_u : \psi(x) = 1\}$ and $h_j : D \to D$ by

$$h_j(x) = f_j(x)/\psi(f_j(x)).$$

It is another easy exercise (again left to the reader) to prove that $\langle h_j \rangle$ satisfies the bounded orbit property if and only if $\langle f_j \rangle$ does. The maps h_j can also be considered maps of $D_1 = \Sigma_\psi$ to itself and $\langle h_j \rangle$ satisfies the bounded orbit property on D_1 if and only if $\langle h_j \rangle$ satisfies the bounded orbit property on D.

Suppose that C is a cone with nonempty interior in a finite-dimensional Banach space X and $f : \mathring{C} = D \to \mathring{C}$ is homogeneous of degree 1 and order-preserving. We can define $f_j = f$ for all $j \geqq 1$ and ask whether $\langle f_j \rangle$ satisfies the bounded orbit property with respect to d. It is a special case of results in § 4 of [25] that $\langle f_j \rangle$ satisfies the bounded orbit property if and only if f has an eigenvector $v \in \mathring{C}$ (so $f(v) = \lambda v$). Note that if f extends continuously to C, it certainly has an eigenvector in $C - \{0\}$, but the question of whether f has an eigenvector in $\mathring{C}$ may be quite subtle, even if $f \in \mathscr{M}_-$. The reader is referred to [25], [26] for further details. Even the simple-looking four-dimensional map $f \in M_-$ in [34] requires some care. For a complete, rigorous analysis see [26].

We will need the following simple geometrical lemma to prove our first weak ergodic theorem.

LEMMA 2.1. *Let C be a cone in a Banach space X and let S be a subset of C. Assume that all elements of S are comparable and that there exists $\rho > 0$ such that $d(x, y) \leqq \rho$ for all $x, y \in S$. If $w \in C$ and $w = x_1 + x_2$ for points $x_1, x_2 \in S$, we have that $x_1 \geqq \lambda w$ or $x_2 \geqq \lambda w$, where $\lambda = \frac{1}{2} \exp(-\rho)$.*

Proof. Suppose that $w = x_1 + x_2$ as above and that $d(x_1, x_2) = \tau \leqq \rho$. It suffices to prove that

$$x_1 \geqq \mu w \quad \text{or} \quad x_2 \geqq \mu w \quad \text{where } \mu = \tfrac{1}{2} \exp(-\tau).$$

It follows easily from the definition of Hilbert's projective metric that

$$d(x_1, w) \leqq \tau \quad \text{and} \quad d(x_2, w) \leqq \tau. \tag{2.1}$$

In fact, we have strict inequality in (2.1) if $\tau > 0$. Formulae (2.1) imply that there exist positive constants α_j and β_j for $j = 1, 2$ such that

$$\alpha_j w \leqq x_j \leqq \beta_j w \quad \text{and} \quad (\beta_j/\alpha_j) \leqq \exp(\tau). \tag{2.2}$$

Assume that $\alpha_j < \mu$ for $j = 1$ and $j = 2$. Then we obtain

$$x_j \leqq \beta_j w = (\alpha_j/\mu)\mu(\beta_j/\alpha_j) w \leqq c_j w, \tag{2.3}$$

where

$$c_j = (\tfrac{1}{2})(\alpha_j/\mu) < \tfrac{1}{2}. \tag{2.4}$$

By adding (2.3) for $j = 1$ and $j = 2$ we obtain

$$w = x_1 + x_2 \leqq (c_1 + c_2) w = cw. \tag{2.5}$$

Since the constant c in (2.5) is less than 1, we have a contradiction, and therefore it must be true that $\alpha_1 \geqq \mu$ or $\alpha_2 \geqq \mu$. □

We will actually use Lemma 2.1 in the following less general version.

LEMMA 2.2. *Let C be a cone in a Banach space X and let $A : X \to X$ be a bounded linear operator such that $A(C) \subset C$ and A has finite projective diameter (so $\Delta(A) < \infty$*

for $\Delta(A)$ as in (1.6)). *Then if $z \in C$ and $z = x + y$ for $x, y \in C$, we have $Ax \geqq \lambda Az$ or $Ay \geqq \lambda Az$, where*

$$\lambda = (\tfrac{1}{2}) \exp(-\Delta(A)).$$

Proof. In the notation of Lemma 2.1, define S by

$$S = \{Ax: x \in C \text{ and } Ax \neq 0\}, \tag{2.6}$$

so the diameter ρ of S with respect to Hilbert's projective metric is $\Delta(A)$. If $Ax = 0$ or $Ay = 0$ for x and y as in the statement of the lemma, the result is obvious. Otherwise, if we define $w = Az$, $x_1 = Ax$, and $x_2 = Ay$, Lemma 2.2 follows immediately from Lemma 2.1. □

With these preliminaries we can establish our first weak ergodic theorem. In reading the statement of Theorem 2.1 below, recall that if A and B are linear maps and C is a cone, we say that $A \leqq B$ (in the partial ordering induced by C) if $A(x) \leqq B(x)$ for all $x \in C$.

THEOREM 2.1. *Let C be a cone with nonempty interior in a Banach space X and for each $j \geqq 1$ let $f_j: \mathring{C} \to \mathring{C}$ be a map that is order-preserving and homogeneous of degree 1. For $m \geqq 1$ let $F_m = f_m f_{m-1} \cdots f_1$ denote the composition of the first m maps f_j and let F_0 denote the identity. Assume that there exist $u \in \mathring{C}$, an integer $p \geqq 1$, a real number $R > 0$, and a sequence of bounded linear operators $A_i: X \to X$, $i \geqq 1$, with the following properties*:

(a) *For every $j \geqq 1$, f_j is continuously Fréchet differentiable on $B_R(F_{j-1}(u))$, where $B_R(y) = \{x \in \mathring{C}: d(x, y) < R\}$.*

(b) *The operator A_i satisfies $A_i(\mathring{C}) \subset (\mathring{C})$ for all $i \geqq 1$.*

(c) *If $g_j = f_{jp} f_{jp-1} \cdots f_{jp-p+1}$ and $G_m = g_m g_{m-1} \cdots g_1 = F_{mp}$, then $g_j'(x) \geqq A_j$ for all $x \in B_R(G_{j-1}(u))$ and all $j \geqq 1$.*

(d) *There exists a positive constant k such that $A_j(G_{j-1}(u)) \geqq k g_j(G_{j-1}(u))$ for $j \geqq 1$.*

If A_j has finite projective diameter, let $\Delta(A_j)$ be the projective diameter as in (1.6), *and otherwise define $\Delta(A_j) = \infty$ and $\exp(-\Delta(A_j)) = 0$. Then if we have that*

$$\lim_{N \to \infty} \sum_{j=1}^{N} \exp(-\Delta(A_j)) = \infty, \tag{2.7}$$

it follows that

$$\lim_{m \to \infty} d(F_m(x), F_m(y)) = 0 \quad \text{for all } x, y \in \mathring{C}. \tag{2.8}$$

Also, if C is normal, (1.13) *is satisfied for all $x, y \in \mathring{C}$. In particular, if $A_j = A$ for all $j \geqq 1$ and $\Delta(A) < \infty$, then* (2.8) *is satisfied.*

Proof. By the triangle inequality it suffices to prove (2.8) for all $x \in \mathring{C}$ and for $y = u$. As has already been noted, each map f_j is nonexpansive with respect to d, so for any $x \in \mathring{C}$, $d(F_m(x), F_m(u))$ is a monotonic decreasing sequence of reals. Thus to prove (2.8) it suffices to prove that

$$\lim_{j \to \infty} d(G_j(x), G_j(u)) = 0 \quad \text{for all } x \in \mathring{C}. \tag{2.9}$$

Fix a number R_1, $0 < R_1 < R$, and suppose we can prove that there exists a sequence of numbers λ_j with $0 < \lambda_j \leqq 1$, such that if $d(y, G_j(u)) \leqq R_1$, $j \geqq 0$, then

$$d(g_{j+1}(y), g_{j+1}(G_j(u))) \leqq \lambda_{j+1} d(y, G_j(u)), \tag{2.10}$$

$$\lim_{N \to \infty} \prod_{j=m}^{N} \lambda_j = 0 \quad \text{for any } m \geqq 1. \tag{2.11}$$

Repeated application of (2.10) and (2.11) then implies that if $d(G_{m-1}(x), G_{m-1}(u)) \leqq R_1$ for some $m \geqq 1$, then

$$d(G_N(x), G_N(u)) \leqq \left(\prod_{j=m}^{N} \lambda_j \right) \quad \text{for all } N \geqq m,$$

which establishes (2.9) in the case $d(x, u) \leqq R_1$.

If $d(y, G_{j-1}(u)) = \rho > R_1$, Proposition 1.9 in [25] implies that there exists y_1 on the line segment connecting y to $G_{j-1}(u)$ such that

$$d(y, y_1) = \rho - R_1 \quad \text{and} \quad d(y_1, G_{j-1}(u)) = R_1. \tag{2.12}$$

Using the nonexpansivity of g_j and (2.10) we obtain from (2.12) that

$$d(g_j(y), G_j(u)) \leqq \rho - R_1 + \lambda_j R_1. \tag{2.13}$$

If $\rho \leqq R_2$, $R_2 > R_1$, we obtain from (2.13) that

$$d(g_j(y), G_j(u)) \leqq \mu_j d(y, G_{j-1}(u)), \tag{2.14}$$

where

$$\mu_j = [R_2 - (1-\lambda_j)R_1]R_2^{-1}. \tag{2.15}$$

Formula (2.14) was proved under the assumption that $R_1 < \rho \leqq R_2$, but because $\mu_j \geqq \lambda_j$, the equation holds for $\rho \leqq R_2$.

If $d(x, u) \leqq R_2$, the nonexpansive property of G_j implies that

$$d(G_j(x), G_j(u)) \leqq R_2 \quad \text{for all } j \geqq 0.$$

Thus by repeated applications of (2.14) we obtain

$$d(G_N(x), G_N(u)) \leqq \left(\prod_{j=1}^{N} \mu_j \right) d(x, u),$$

so (2.9) will follow if we can prove that

$$\lim_{N\to\infty} \left(\prod_{j=1}^{N} \mu_j \right) = 0. \tag{2.16}$$

If $\lambda_j \leqq \frac{1}{2}$ for infinitely many indices j, we can easily see that there exists a constant $c < 1$ such that $0 \leqq \mu_j \leqq c$ for infinitely many indices c and (2.16) will be satisfied. Thus we can assume that $\frac{1}{2} < \lambda_j \leqq 1$ for $j \geqq m$, so $\frac{1}{2} < \mu_j \leqq 1$ for $j \geqq m$. Under these conditions it is well known and easily checked that

$$\lim_{N\to\infty} \prod_{j=m}^{N} \lambda_j = 0 \Leftrightarrow \sum_{j=m}^{\infty} -\log(\lambda_j) = \infty \Leftrightarrow \sum_{j=m}^{\infty} (1-\lambda_j) = \infty,$$

$$\lim_{N\to\infty} \prod_{j=m}^{N} \mu_j = 0 \Leftrightarrow \sum_{j=m}^{\infty} -\log(\mu_j) = \infty \Leftrightarrow \sum_{j=m}^{\infty} (1-\mu_j) = \infty.$$

Because $(1-\mu_j) = (R_1/R_2)(1-\lambda_j)$, the equations above imply that (2.16) is satisfied.

Thus it suffices to prove that (2.10) and (2.11) can be satisfied. For a fixed $\psi \in C^* - \{0\}$ it suffices (by homogeneity) to define $u_j = G_j(u)/\psi(G_j(u))$ and to prove that

$$d(g_{j+1}(u_j), g_{j+1}(y)) \leqq \lambda_{j+1} d(y, u_j)$$

for all y such that $\psi(y) = 1$ and $d(y, u_j) \leqq R_1 < R$. Recall (see Lemma 4.1 in [25] or argue directly) that

$$V_{R_1}(u_j) = \{y: d(y, u_j) \leqq R_1\}$$

is convex. For notational convenience, define $G_{j+1} = g$, $A_{j+1} = A$, and $u_j = v$ and let α and β be positive numbers such that

$$\alpha v \leqq y \leqq \beta v \quad \text{and} \quad \log(\beta/\alpha) = \rho = d(y, v).$$

The normalization $\psi(y) = \psi(v) = 1$ implies that $\alpha \leqq 1$ and $\beta \geqq 1$. If we define

$$w_t = (1-t)(\alpha v) + ty \quad \text{and} \quad z_t = (1-t)y + t(\beta v), \quad 0 \leqq t \leqq 1,$$

we obtain

$$g(y) - g(\alpha v) = \int_0^1 g'(w_t)(y - \alpha v)\, dt,$$

$$g(\beta v) - g(y) = \int_0^1 g'(z_t)(\beta v - y)\, dt.$$

If we recall that $g'(z_t) \geqq A$, we obtain from the preceding two equations that

$$\alpha g(v) + A(y - \alpha v) \leqq g(y) \leqq \beta g(v) - A(\beta v - y). \tag{2.17}$$

Note that

$$A(y - \alpha v) + A(\beta v - y) = (\beta - \alpha)Av,$$

so Lemma 2.2 implies that there exists a positive constant $\gamma = \gamma_{j+1} = (\frac{1}{2}) \exp(-\Delta(A_{j+1}))$ such that

$$A(y - \alpha v) \geqq \gamma(\beta - \alpha)Av \quad \text{or} \quad A(\beta v - y) \geqq \gamma(\beta - \alpha)Av.$$

For definiteness we assume that

$$A(y - \alpha v) \geqq \gamma(\beta - \alpha)Av. \tag{2.18}$$

The proof in the other case is essentially the same.

Next, remember that we assume the existence of a positive constant k, independent of $j \geqq 0$, such that

$$A(v) = A_{j+1}(u_j) \geqq kg(v) = kg_{j+1}(u_j). \tag{2.19}$$

If we use (2.17)-(2.19) we obtain

$$\alpha g(v) + k\gamma(\beta - \alpha)g(v) \leqq g(y) \leqq \beta g(v), \tag{2.20}$$

where $\log(\beta/\alpha) = d(y, v)$ and $\gamma = (\frac{1}{2}) \exp(-\Delta(A_{j+1}))$. Formula (2.20) implies that $k\gamma \leqq 1$ and

$$d(g(v), g(y)) \leqq \log\left(\frac{\beta}{\alpha + k\gamma(\beta - \alpha)}\right). \tag{2.21}$$

If we define $s = \beta/\alpha$, with $1 \leqq s \leqq \exp(R_1)$, and recall that $d(v, y) = \log(s)$, we obtain from (2.21) that $d(g(v), g(y)) = d(g_{j+1}(u_j), g_{j+1}(y)) \leqq \lambda_{j+1} d(u_j, y)$, where

$$\lambda_{j+1} = \sup_{1 < s \leqq \exp(R_1)} \frac{\varphi_1(s)}{\varphi_2(s)}, \tag{2.22}$$

with $\varphi_1(s) = \log[s(1 + k\gamma(s-1))^{-1}]$ and $\varphi_2(s) = \log(s)$.

Because $\varphi_j(1) = 0$, the generalized mean value theorem implies that for each s, $1 < s \leqq \exp(R_1)$, there exists σ with $1 < \sigma < s$ such that

$$\varphi_1(s)/\varphi_2(s) = \varphi_1'(\sigma)/\varphi_2'(\sigma) = (1 - k\gamma)(1 + k\gamma(\sigma - 1))^{-1}, \tag{2.23}$$

so

$$\lambda_{j+1} = 1 - \left(\frac{k}{2}\right) \exp\left(-\Delta(A_{j+1})\right). \tag{2.24}$$

The same sort of argument that we have used already proves that

$$\lim_{N\to\infty} \prod_{j=m}^{N} \lambda_j = 0 \quad \text{for every } m \geqq 1$$

if and only if (2.7) is satisfied.

This completes the proof of (2.8); the remaining assertions of the theorem are straightforward and left to the reader. □

Remark 2.1. Let hypotheses and notation be as in Theorem 2.1. However, do not assume condition (d), and suppose that $\langle g_j \rangle$ satisfies the bounded orbit property. Then condition (d) is equivalent to the following condition (d′):

(d′) There exists a positive constant k_1 such that

$$A_j(u) \geqq k_1 g_j(u) \quad \text{for all } j \geqq 1.$$

We will prove that (d′) implies (d); the opposite implication is proved similarly. If u_j is as defined in the proof of Theorem 2.1, the bounded orbit hypothesis implies that there exist positive constants c_j and d_j, $j \geqq 1$, such that

$$c_j u_0 \leqq u_j \leqq d_j u_0 \quad \text{and} \quad d_j / c_j \leqq M,$$

where M is a constant independent of j. It follows that

$$c_j g_{j+1}(u_0) \leqq g_{j+1}(u_j) \leqq d_j g_{j+1}(u_0),$$

$$c_j A_{j+1}(u_0) \leqq A_{j+1}(u_j) \leqq d_j A_{j+1}(u_0).$$

If we use these inequalities and hypothesis (d′) we find that

$$g_{j+1}(u_j) \leqq d_j g_{j+1}(u_0) \leqq k_1^{-1} d_j A_{j+1}(u_0) \leqq k_1^{-1} \left(\frac{d_j}{c_j}\right) A_{j+1}(u_j)$$

$$\leqq k_1^{-1} M A_{j+1}(u_j),$$

which is equivalent to hypothesis (d).

It may be unclear how we can expect to find operators A_j such as those in Theorem 2.1. The next corollary shows that under mild assumptions a scalar multiple of $g_j'(G_{j-1}(u))$ can serve as A_j.

COROLLARY 2.1. *Let C be a cone with nonempty interior in a Banach space X, and let $f_j : \mathring{C} \to \mathring{C}$, $j \geqq 1$, be a sequence of maps that are order-preserving and homogeneous of degree* 1. *For a fixed $p \geqq 1$, let g_j and G_m be as defined in Theorem* 2.1, *and assume that there exist $r > 0$ and $u \in \mathring{C}$ such that f_j is C^1 on $B_r(F_{j-1}(u))$ for all $j \geqq 1$, where $B_r(x) = \{y \in \mathring{C} : d(y, x) < r\}$. Define $u_j = G_j(u)/\|G_j(u)\|$ and assume that there exist positive constants c and $\rho \leqq r$ such that, if $d(x, u_{j-1}) < \rho$ and $j \geqq 1$, then*

$$g_j'(x) \geqq c g_j'(u_{j-1}). \tag{2.25}$$

Finally, assume that the operators $B_j = g_j'(G_{j-1}(u)) = g_j'(u_{j-1})$ satisfy

$$\sum_{j=1}^{\infty} \exp\left(-\Delta(B_j)\right) = \infty, \tag{2.26}$$

where $\Delta(B_j)$ *is given by* (1.6) *if* B_j *has finite projective diameter, and that they satisfy* $\exp(-\Delta(B_j)) = 0$ *otherwise. Then it follows that*

$$\lim_{n\to\infty} d(F_n(x), F_n(y)) = 0 \quad \text{for all } x, y \in \mathring{C}.$$

If C *is normal,* (1.13) *is also satisfied for all* $x, y \in \mathring{C}$.

Proof. Define $A_j = cB_j$ for c as in (2.25). It suffices to prove that the hypotheses of Theorem 2.1 are satisfied. Formula (2.25) implies (taking $\rho = R$) that hypothesis (c) of Theorem 2.1 is satisfied.

Because g_i is order-preserving, B_i and $A_i = cB_i$ are also order-preserving, so $A_i(C) \subset C$. The homogeneity of g_i implies that

$$g_i(u_{i-1}) = B_i(u_{i-1}) \in \mathring{C}, \tag{2.27}$$

and using (2.27) and the order-preserving property of A_i we conclude that $A_i(\mathring{C}) \subset \mathring{C}$. Thus hypothesis (b) of Theorem 2.1 is satisfied. Hypothesis (d) of Theorem 2.1 also follows directly from (2.27). Finally, because $\Delta(cB_i) = \Delta(B_i)$, (2.7) is equivalent to (2.26). □

The homogeneity of the functions f_j in Theorem 2.1 plays less of a role than it might at first seem to. We illustrate this by stating a result that follows by essentially the same argument as Theorem 2.1. First, we need a lemma proved by Potter (see [29]) in the case where the function g is defined, order-preserving, and subhomogeneous on *all* of $\mathring{C}$.

LEMMA 2.3. (Compare [29].) *Let* C *be a cone with nonempty interior in a Banach space* X, *and for* $\psi \in C^* - \{0\}$ *and* $\lambda > 1$ *define* $\Sigma = \{x \in \mathring{C}: \psi(x) = 1\}$ *and* $D = \{x \in \mathring{C}: \lambda^{-1} < \psi(x) < \lambda\}$. *Assume that* $f: D \to \mathring{C}$ *is order-preserving and subhomogeneous on* D. *Then* $f|\Sigma$ *is nonexpansive with respect to Hilbert's projective metric* d, *so* $d(f(x), f(y)) \leqq d(x, y)$ *for all* $x, y \in \Sigma$.

Proof. If x_0 and x_1 are points in Σ with $d(x_0, x_1) = \rho$, define $x_s = (1-s)x_0 + sx_1$. For a given integer $n > 1$, it follows from Proposition 1.9 in [25] that there are real numbers s_j, with $s_j < s_{j+1}$ for $0 \leqq j < n$ and $s_0 = 0$ and $s_n = 1$ and (writing $y_j = x_{s_j}$)

$$d(y_j, y_{j+1}) = \rho n^{-1} \quad \text{for } 0 \leqq j < n.$$

Note that in general $s_j \neq jn^{-1}$.

Choose n so large that $\rho n^{-1} < \lambda$. It suffices to prove that

$$d(f(y_j), f(y_{j+1})) \leqq d(y_j, y_{j+1}) = \rho n^{-1}, \tag{2.28}$$

for then the triangle inequality gives

$$d(f(x_0), f(x_1)) \leqq \sum_{j=0}^{n-1} d(f(y_j), f(y_{j+1})) \leqq n(\rho n^{-1}) = d(x_0, x_1).$$

For a fixed j select numbers α and β so that

$$\alpha y_j \leqq y_{j+1} \leqq \beta y_j \quad \text{and} \quad \log\left(\frac{\beta}{\alpha}\right) = \rho n^{-1}. \tag{2.29}$$

Because $\psi(y_j) = \psi(y_{j+1}) = 1$, we easily obtain from (2.29) that $0 < \alpha \leqq 1 \leqq \beta$ and $\alpha^{-1} \leqq \rho n^{-1} < \lambda$ and $\beta < \rho n^{-1} < \lambda$. It follows that the points αy_j, $\beta^{-1} y_{j+1}$, y_j, and y_{j+1} all lie in D. By using the subhomogeneity and the order-preserving property of f on D we find that

$$\alpha f(y_j) \leqq f(\alpha y_j) \leqq f(y_{j+1}) \quad \text{and} \quad \beta^{-1} f(y_{j+1}) \leqq f(\beta^{-1} y_{j+1}) \leqq f(y_j),$$

so

$$d(f(y_j), f(y_{j+1})) \leqq \log\left(\frac{\beta}{\alpha}\right).$$

□

We can now give a version of Theorem 2.1 for order-preserving, subhomogeneous operators. Once we know Lemma 2.3, the proof of the next theorem follows by essentially the same argument as in Theorem 2.1 and is left to the reader.

THEOREM 2.2. *Let C be a cone with nonempty interior in a Banach space X and for fixed $\psi \in C^* - \{0\}$ and $\lambda > 1$ define $D = \{x \in \mathring{C}: \lambda^{-1} < \psi(x) < \lambda\}$ and $\Sigma = \{x \in \mathring{C}: \psi(x) = 1\}$. Suppose that $g_j: D \to \mathring{C}$, $j \geqq 1$, is a sequence of order-preserving, subhomogeneous maps and define*

$$h_j(x) = g_j(x)/\psi(g_j(x)),$$

with $H_m = h_m h_{m-1} \cdots h_1$ and H_0 equal to the identity. For some $u \in \Sigma$ define $u_j = H_j(u)$ and assume that there exists $R > 0$ and a sequence of bounded linear operators A_j, $j \geqq 1$, such that:

(a) *g_j is continuously Fréchet differentiable on $B_R(u_{j-1}) \cap D$ for all $j \geqq 1$, where $B_R(y) = \{x \in \mathring{C}: d(x, y) < R\}$.*

(b) *$A_i(\mathring{C}) \subset \mathring{C}$ for all $i \geqq 1$ and $g_i'(x) \geqq A_i$ for all $x \in B_R(u_{i-1}) \cap D$ and $i \geqq 1$.*

(c) *There exists a positive constant k such that $A_j(u_{j-1}) \geqq k g_j(u_{j-1})$ for all $j \geqq 1$.*

If we have

$$\lim_{N \to \infty} \sum_{j=1}^{N} \exp(-\Delta(A_j)) = \infty,$$

where $\Delta(A_j)$ is given by (1.6), *then it follows that*

$$\lim_{n \to \infty} d(H_n(x), H_n(y)) = 0 \quad \text{for all } x, y \in \Sigma,$$

and

$$\lim_{n \to \infty} \|H_n(x) - H_n(y)\| = 0 \quad \text{for all } x, y \in \Sigma$$

if C is normal.

Remark 2.2. Note that in the statement of Theorem 2.2 there is no integer p analogous to the one in the statement of Theorem 2.1. If we assume that the g_j satisfy a "ray-preserving property" as in [16], then

$$\{g_j(tx): t > 0\} \subset \{s g_j(x): s > 0\} \quad \text{for } j \geqq 1 \quad \text{and} \quad x \in \Sigma.$$

Then if $G_m = g_m g_{m-1} \cdots g_1$, we can verify (see [16]) that

$$G_m(x)/\psi(G_m(x)) = H_m(x).$$

Using this fact we can then give a version of Theorem 2.2 that directly generalizes Theorem 2.1. Without the ray-preserving property there seems to be no necessary connection between G_m and H_m.

However, the assumption of the ray-preserving property for order-preserving, subhomogeneous operators can be restrictive. To see this, suppose that $f_j: \mathring{C} \to \mathring{C}$, $j = 1, 2$, is order-preserving and $f_j(tx) = t^{\lambda_j} f_j(x)$ for $0 < t \leqq 1$ and $x \in \mathring{C}$, where $0 < \lambda_j \leqq 1$. Then f_j is subhomogeneous, order-preserving, and ray-preserving. However, $f = f_1 + f_2$ is subhomogeneous and order-preserving, but not ray-preserving unless $\lambda_1 = \lambda_2$.

Remark 2.3. In [16], Fujimoto and Krause have obtained weak ergodic theorems for ray-preserving maps of the standard cone $K = \{x \in \mathbb{R}^n: x_i \geqq 0 \text{ for } 1 \leqq i \leqq n\}$ into itself. If $\Gamma = \{x \in K: \psi(x) = 1\}$ ($\psi \in K^* - \{0\}$) and $f_j: K \to K$, $j \geqq 1$, is a sequence of maps for which we want to establish a weak ergodic theorem, then the assumption of "uniform pointwise boundedness" in Theorem 4 of [16] (assuming, for simplicity that $r = 1$ in the statement of that theorem) implies that there exist $a, b \in \mathring{K}$ such that

$$a \leqq f_j(x)/\psi(f_j(x)) = h_j(x) \leqq b \quad \text{for all } x \in \Gamma. \tag{2.30}$$

Formula (2.30) (or its analogue for $r \geqq 1$) implies that $h_j : \Gamma \cap \mathring{K} \to \Gamma \cap \mathring{K}$, $j \geqq 1$, satisfies the bounded orbit property with respect to d, but uniform pointwise boundedness represents a less general condition than the bounded orbit property does. In particular, if $f_j \in \mathcal{M}$ for $j \geqq 1$, many examples of interest do not satisfy (2.30) or even the condition that $f_j(K - \{0\}) \subset \mathring{K}$, but they may satisfy the bounded orbit property. Of course (see § 4 of [16]), there are also many examples where the uniform pointwise boundedness assumption is satisfied.

Remark 2.4. The reader will note that the bounded orbit property is not assumed in Theorem 2.1, Corollary 2.1, or Theorem 2.2. Nevertheless, the bounded orbit property plays a crucial role: to verify the hypotheses of Corollary 2.1 or Theorem 2.2 for examples of interest, we will typically have to verify the bounded orbit property.

In the framework of Theorem 2.1 it is natural to ask if a stronger conclusion can be obtained. Does there exist $v \in \mathring{C}$ such that

$$\lim_{n\to\infty} d(F_n(x), v) = 0 \quad \text{and} \quad \lim_{n\to\infty} \| F_n(x)/\|F_n(x)\| - v\| = 0 \quad \text{for all } x \in \mathring{C}?$$

A similar question can be asked for Theorem 2.2. Such results are called "strong ergodic theorems" in [11]. As we now show, such a theorem can be derived easily from Theorem 2.1.

THEOREM 2.3. *Let C be a cone with nonempty interior in a Banach space X and $f_j : \mathring{C} \to \mathring{C}$, $j \geqq 1$, a sequence of maps that are homogeneous of degree 1 and order-preserving. Assume that there exists $v \in \mathring{C}$, $\|v\| = 1$, such that*

$$\lim_{j\to\infty} d(f_j(v), v) = 0, \tag{2.31}$$

where d denotes Hilbert's projective metric. Assume there exists $R > 0$, an integer $p \geqq 1$, and a sequence of bounded linear operators A_j, $j \geqq 1$, with the following properties:

(a) *If g_j is defined as in Theorem 2.1, g_j is continuously Fréchet differentiable on $B_R(v) = \{x \in \mathring{C} : d(x, v) < R\}$ for all $j \geqq 1$.*

(b) *For all $x \in B_R(v)$ and all $j \geqq 1$ we have $g_j'(x) \geqq A_j$.*

(c) *The operators A_j satisfy $A_j(\mathring{C}) \subset \mathring{C}$, A_j has finite projective diameter $\Delta(A_j)$, and*

$$\sup\{\Delta(A_j) : j \geqq 1\} < \infty. \tag{2.32}$$

(d) *There exists a positive constant k such that $A_j(v) \geqq k g_j(v)$ for all $j \geqq 1$. Then it follows that $\lim_{n\to\infty} d(F_n(x), v) = 0$ for all $x \in \mathring{C}$, and $\lim_{n\to\infty} \|F_n(x)/\|F_n(x)\| - v\| = 0$ if C is normal.*

Proof. By using the triangle inequality and the nonexpansiveness of f_j we can verify that

$$d(g_j(v), v) \leqq \sum_{i=jp-p+1}^{jp} d(f_i(v), v),$$

so

$$\lim_{j\to\infty} d(g_j(v), v) = 0.$$

We claim that to prove the theorem it suffices to prove that

$$\lim_{n\to\infty} d(G_n x, v) = \lim_{n\to\infty} d(F_{np} x, v) = 0 \quad \text{for all } x \in \mathring{C}. \tag{2.33}$$

To see this, note that the triangle inequality and the nonexpansiveness of the f_j imply that for $np < m < np + p$,

$$d(F_m x, v) \leqq d(F_{np} x, v) + \sum_{j=0}^{m-np-1} d(f_{m-j} v, v). \tag{2.34}$$

Formulae (2.31), (2.33), and (2.34) imply that

$$\lim_{m\to\infty} d(F_m x, v) = 0 \quad \text{for all } x \in \mathring{C}.$$

Thus it suffices to work with g_j and G_j.

Select a fixed number R_1, $0 < R_1 < R$, and define $V_{R_1}(v) = \{z: d(z, v) \leqq R_1\}$. Essentially the same argument used in the proof of Theorem 2.1 actually shows that for all $x, y \in V_{R_1}(v)$,

$$d(g_j(x), g_j(y)) \leqq c_j d(x, y),$$

where $c_j = [1 - (k/2)\, e^{-R_1} \exp(-\Delta(A_j))]$. (Note the extra factor e^{-R_1} in c_j; this is unnecessary if x or y equals v.) Because we assume that $\Delta(A_j)$ is bounded above, we have

$$\sup\{c_j: j \geqq 1\} = c < 1.$$

Now select any number ε, $0 < \varepsilon < R_1$, and suppose that $x \in V_\varepsilon(v)$, so $d(x, v) \leqq \varepsilon$. Then we obtain

$$\begin{aligned} d(g_j(x), v) &\leqq d(g_j(x), g_j(v)) + d(g_j(v), v) \\ &\leqq c\varepsilon + d(g_j(v), v). \end{aligned}$$

It follows that there exists an integer $m = m(\varepsilon)$ such that

$$g_j(V_\varepsilon(v)) \subset V_\varepsilon(v) \quad \text{for } j \geqq m(\varepsilon).$$

We now apply Theorem 2.1. For $m = m(\varepsilon)$ as above, define $\varphi_i(x) = g_{m+i}(x)$ and $\Phi_j = \varphi_j \varphi_{j-1} \cdots \varphi_1$. The sequence φ_j, $j \geqq 1$, satisfies the bounded orbit property; in fact, $\Phi_j(v) \in V_\varepsilon(v)$ for all $j \geqq 1$. It is also easy to check that all the hypotheses of Theorem 2.1 are satisfied, so

$$\lim_{j\to\infty} d(\Phi_j(y), \Phi_j(v)) = 0 \quad \text{for all } y \in \mathring{C}.$$

Taking $y = G_m(x)$ for $x \in \mathring{C}$, we conclude that

$$\lim_{j\to\infty} d(G_{j+m}(x), \Phi_j(v)) = 0 \quad \text{for all } x \in \mathring{C}.$$

Since $\Phi_j(v) \in V_\varepsilon(v)$ for all $j \geqq 1$, we conclude that for any fixed $x \in \mathring{C}$ there exists n such that

$$d(G_j(x), v) < 2\varepsilon \quad \text{for all } j \geqq n.$$

Since $\varepsilon > 0$ can be taken as small as desired, the proof is complete. □

Remark 2.5. Suppose that $f: \mathring{C} \to \mathring{C}$ is a map and $f(v) = \lambda v$ for some $v \in \mathring{C}$. Assume that $f_j: \mathring{C} \to \mathring{C}$, $j \geqq 1$, is a sequence of maps as in Theorem 2.3 and that for every $x \in \mathring{C}$ we have

$$\lim_{j\to\infty} d(f_j(x), f(x)) = 0.$$

Then it is certain that $\lim_{j\to\infty} d(f_j(v), v) = 0$, which provides a situation where we can find a vector v as in Theorem 2.3. However, such a vector v may well exist even if the functions f_j do not converge to a function f.

Remark 2.6. Essentially the same argument as in Theorem 2.3 can be extended, as in Theorem 2.2, to the case of order-preserving maps that are subhomogeneous. Details are left to the reader.

Under the hypotheses of Theorem 2.1, it is natural to ask if, for every given $x \in \mathring{C}$, there exists a positive number $\gamma = \gamma(x)$ such that

$$\lim_{n\to\infty} \|F_n(x) - \gamma F_n(u)\| = 0.$$

If $f_j = f$ for all $j \geqq 1$ and $f(u) = u$, this question is considered at length in § 3 of [25]. Before considering the general case it is convenient to prove a simple lemma.

LEMMA 2.4. *Let C be a normal cone with nonempty interior in a Banach space X. If $u \in \mathring{C}$ and R, c_1, and c_2 are positive reals, define V by*

$$V = \{x \in \mathring{C}: d(x, u) \leqq R \text{ and } c_1 \leqq \|x\| \leqq c_2\},$$

where d denotes Hilbert's projective metric. Then there exists $\rho > 0$ so that, for every $x \in V$,

$$\{z \in X: \|z - x\| < \rho\} \subset \mathring{C}.$$

Proof. If $x \in V$, there exist positive numbers α and β such that

$$\alpha u \leqq x \leqq \beta u \quad \text{and} \quad \beta\alpha^{-1} \leqq e^R. \tag{2.35}$$

By the definition of normality there exists a constant M so that if $0 \leqq x \leqq y$ then

$$\|x\| \leqq M\|y\|. \tag{2.36}$$

Combining (2.35) and (2.36) gives

$$\alpha\|u\| \leqq M\|x\| \leqq Mc_2 \quad \text{and} \quad c_1 \leqq \|x\| \leqq M\beta\|u\|. \tag{2.37}$$

Formulae (2.35) and (2.37) imply that

$$(c_1/M\|u\|)\, e^{-R} \leqq \alpha \quad \text{and} \quad \beta \leqq (Mc_2/\|u\|)\, e^R.$$

Thus there exists a number $\gamma \geqq 1$, independent of $x \in V$, such that

$$\gamma^{-1}u \leqq x \leqq \gamma u. \tag{2.38}$$

Because $\gamma^{-1}u \in \mathring{C}$, there exists $\rho > 0$ such that $\gamma^{-1}u + z \in \mathring{C}$ for all $z \in X$ with $\|z\| < \rho$, and (2.38) then implies that $x + z \in \mathring{C}$ for all $x \in V$ and all z with $\|z\| < \rho$. □

If C is a cone with nonempty interior in a Banach space X, $x \in \mathring{C}$ and $\{y: \|y - x\| < \rho\} \subset \mathring{C}$, then we can easily prove that for $\|y - x\| < \rho$ we have

$$(\rho - \|y - x\|)\rho^{-1}x \leqq y \leqq (\rho + \|y - x\|)\rho^{-1}x. \tag{2.39}$$

THEOREM 2.4. *Let notation and assumptions be as in Theorem 2.1. Assume also that $\langle f_j \rangle$ satisfies the bounded orbit property, that C is normal, and that*

$$\sup\{\|F_m(u)\|: m \geqq 0\} < \infty \quad \textit{and} \quad \inf\{\|F_m(u)\|: m \geqq 0\} > 0. \tag{2.40}$$

Then for every $x \in \mathring{C}$ there exists a positive number $\gamma = \gamma(x)$ such that

$$\lim_{m\to\infty} \|F_m(x) - \gamma F_m(u)\| = 0. \tag{2.41}$$

The map $x \to \gamma(x)$ is homogeneous of degree 1, order-preserving, and continuous.

Proof. It is known (see [32]) that there exists an equivalent norm on X whose restriction to C is order-preserving. Thus we can assume that if $0 \leqq x \leqq y$, then $\|x\| \leqq \|y\|$.

For a fixed $x \in \mathring{C}$, define numbers α_n and β_n by

$$\alpha_n = m(F_n(x)/F_n(u)) \quad \text{and} \quad \beta_n = M(F_n(x)/F_n(u)),$$

so

$$\alpha_n F_n(u) \leqq F_n(x) \leqq \beta_n F_n(u). \tag{2.42}$$

Applying f_{n+1} to (2.42) we find that

$$\alpha_n F_{n+1}(u) \leqq F_{n+1}(x) \leqq \beta_n F_{n+1}(u),$$

and we conclude from this inequality that

$$\alpha_n \leqq \alpha_{n+1}, \quad \beta_{n+1} \leqq \beta_n, \quad \alpha_n \leqq \beta_n \quad \text{for all } n, \tag{2.43}$$

so $\lim_{n\to\infty} \alpha_n = \alpha$ and $\lim_{n\to\infty} \beta_n = \beta$.

If we can prove that $\alpha = \beta$, we can define $\gamma(x) = \beta$ and (by normality of the cone) it is evident from (2.42) that (2.41) is satisfied. Furthermore, the homogeneity and order-preserving properties of $\gamma(x)$ follow immediately from the formula

$$\gamma(x) = \lim_{n\to\infty} M(F_n(x)/F_n(u)).$$

Because γ is homogeneous of degree 1 and order-preserving, we can easily derive from (2.39) that γ is continuous.

Theorem 2.1 implies that

$$\lim_{n\to\infty} \|F_n(x)\|F_n(x)\|^{-1} - F_n(u)\|F_n(u)\|^{-1}\| = 0, \tag{2.44}$$

and (2.42) gives

$$\alpha_n \leqq \gamma_n = \|F_n(x)\|/\|F_n(u)\| \leqq \beta_n. \tag{2.45}$$

Formulae (2.40) and (2.45) imply that $\|F_n(x)\|$ is bounded above and below by positive reals. Using the fact that $\|F_n(x)\|$ is bounded above, we obtain from (2.44)

$$\lim_{n\to\infty} \|F_n(x) - \gamma_n F_n(u)\| = \lim_{n\to\infty} \varepsilon_n = 0. \tag{2.46}$$

We now use Lemma 2.4. The bounded orbit property implies that there exists $R > 0$ such that

$$d(\gamma_n F_n(u), u) = d(F_n(u), u) \leqq R \quad \text{for all } n \geqq 1.$$

Furthermore, as already noted, there exist positive reals c_1 and c_2 such that

$$c_1 \leqq \gamma_n \|F_n(u)\| = \|F_n(x)\| \leqq c_2.$$

Thus Lemma 2.3 implies that there exists $\rho > 0$ such that if $\|z - \gamma_n F_n(u)\| < \rho$ for some $n \geqq 0$, then $z \in \mathring{C}$. If ε_n is defined as in (2.46), then (2.39) implies that if $\varepsilon_n < \rho$ for $n \geqq N$, then for $n \geqq N$,

$$[(\rho - \varepsilon_n)\rho^{-1}]\gamma_n F_n(u) \leqq F_n(x) \leqq [(\rho + \varepsilon_n)\rho^{-1}]\gamma_n F_n(u). \tag{2.47}$$

It follows that

$$0 \leqq \beta_n - \alpha_n \leqq (\rho + \varepsilon_n)\rho^{-1}\gamma_n - (\rho - \varepsilon_n)\rho^{-1}\gamma_n = 2\varepsilon_n \rho^{-1}\gamma_n,$$

and we conclude (using (2.46) and the fact γ_n is bounded) that

$$\lim_{n\to\infty} (\beta_n - \alpha_n) = 0. \qquad \square$$

As an easy corollary of Theorem 2.4 we mention the following result, a slightly weaker version of which has been proved by Cohen in [12]. Of course, in this simple situation it is also possible to give an elementary, direct proof, so the following corollary is meant only as an illustration of Theorem 2.4.

COROLLARY 2.2. (Compare [12].) *Let* $K = \{(x_1, x_2) \in \mathbb{R}^2 : x_i \geqq 0 \text{ for } i = 1, 2\}$. *Let* λ_j, $j \geqq 1$, *be a sequence of real numbers such that* $0 \leqq \lambda_j \leqq 1$, *and for each* $j \geqq 1$ *define*

$$f_j : \mathring{K} \to \mathring{K} \quad \text{by} \quad f_j(x_1, x_2) = ((1-\lambda_j)x_1 + \lambda_j x_2, x_1^{1-\lambda_j} x_2^{\lambda_j})$$

and define $F_n = f_n f_{n-1} \cdots f_1$. *Then for each* $x = (x_1, x_2) \in \mathring{K}$, *there exists* $\gamma = \gamma(x) > 0$ *such that*

$$\lim_{n\to\infty} F_n(x) = (\gamma, \gamma).$$

Proof. Let $u = (1, 1)$ so $f_j(u) = u$ for all $j \geqq 1$. Thus f_j, $j \geqq 1$, satisfies the bounded orbit property in Theorem 2.1 and (2.40) in Theorem 2.4. The functions f_j are clearly all order-preserving, homogeneous of degree 1, and C^1 on $\mathring{K}$. For a fixed $R > 0$, define

$$A_j = e^{-R} \begin{pmatrix} 1-\lambda_j & \lambda_j \\ 1-\lambda_j & \lambda_j \end{pmatrix}.$$

It is easy to check that $f_j'(x) \geqq A_j$ for all x with $d(x, u) \leqq R$ and

$$A_j(u) \geqq e^{-R} f_j(u) = e^{-R} u.$$

Also, A_j has one-dimensional range, so $\Delta(A_j) = 0$. If we take $p = 1$ (in the notation of Theorem 2.1), we thus see that all hypotheses of Theorems 2.1 and 2.4 are satisfied, so the conclusion of the corollary follows from Theorem 2.4. □

The preceding theorems typically make some assumption of differentiability. These assumptions are motivated by the applications we have in mind and can certainly be weakened. The hypotheses of Theorem 2.1 represent only a convenient way to obtain the estimates in (2.10) and (2.11). To illustrate this point we mention the following theorem, whose proof is essentially the first part of the proof of Theorem 2.1. Details are left to the reader.

THEOREM 2.5. *Let* C *be a cone with nonempty interior in a Banach space* X. *Let* Σ *be a subset of* $\mathring{C}$ *and* $S = \{y \in \mathring{C} : \|y\| = 1\}$ *and assume that for each* $y \in S$ *there is a unique positive number* $\lambda = \lambda(y)$ *such that* $\lambda y \in \Sigma$. *Suppose that the* $h_j : \Sigma \to \Sigma$, $j \geqq 1$, *is a sequence of maps and that there exists* $u \in \Sigma$ *such that if* $H_m = h_m h_{m-1} \cdots h_1$ (*the composition of the first* m *functions* h_j) *and* H_0 *denotes the identity, then*

$$d(h_m(x), H_m(u)) \leqq d(x, H_{m-1}(u)) \quad \text{for all } x \in \Sigma \quad \text{and} \quad m \geqq 1.$$

(*Here* d *denotes Hilbert's projective metric.*) *In addition, assume that there exist* $R > 0$ *and a sequence of reals* λ_j, $j \geqq 1$, *such that* $0 \leqq \lambda_j \leqq 1$ *for all* j,

$$d(h_m(x), H_m(u)) \leqq \lambda_m d(x, H_{m-1}(u)) \quad \text{for all } x \in \Sigma \text{ with } d(x, H_{m-1}(u)) \leqq R, \text{ and}$$

$$\lim_{N\to\infty} \prod_{j=m}^{N} \lambda_j = 0 \quad \text{for all } m \geqq 1.$$

It then follows that for every $x \in \Sigma$,

$$\lim_{n\to\infty} d(H_n(x), H_n(u)) = 0.$$

3. Applications: verifying the bounded orbit property. In this section we show how the results of § 2 can be applied in the case where $f_j \in M$ for all $j \geqq 1$, M being the class defined in § 1. We shall show that the main difficulty lies in verifying the bounded orbit property with respect to d. If $f_j \in M_+$ for all $j \geqq 1$, we will give reasonably general conditions ensuring that the bounded orbit property is satisfied. As we have already noted in § 2, if $f_j \in M_-$ for all $j \geqq 1$, verifying the bounded orbit property can be a difficult problem even when $f_j = f$ for all $j \geqq 1$.

We begin with some needed notation and definitions. We will always denote by K the standard cone in $\mathbb{R}^n$,

$$K = \{x \in \mathbb{R}^n : x_i \geqq 0 \text{ for } 1 \leqq i \leqq n\}. \tag{3.1}$$

If $f_j \in M$ is a sequence of functions for $j \geqq 1$, we will denote by f_{ji}, $1 \leqq i \leqq n$, the ith component of the function $f_j : \mathring{K} \to \mathring{K}$. By definition of M, there is a finite collection Γ_{ji} of ordered pairs (r, σ), r a real number and σ a probability vector, and positive real numbers $c_{jir\sigma}$ for $(r, \sigma) \in \Gamma_{ji}$, such that

$$f_{ji}(x) = \sum_{(r,\sigma)\in\Gamma_{ji}} c_{jir\sigma} M_{r\sigma}(x). \tag{3.2}$$

Because $(x^r)^{1/r} = x$, the set Γ_{ji} described above may not be uniquely determined by the function f_{ji}. However, we will need some control of the size of the numbers r, which appear in (3.2). Thus we make the following definitions. Suppose that $\phi \in M$ so ϕ_i, the ith component of ϕ, can be written

$$\phi_i(x) = \sum_{(r,\sigma)\in G_i} c_{ir\sigma} M_{r\sigma}(x), \tag{3.3}$$

where G_i is a finite collection of ordered pairs (r, σ), r a real number and σ a probability vector, and $c_{ir\sigma} > 0$. If $\phi \in M_+$, the sets G_i can be chosen so that $r \geqq 0$ for all $(r, \sigma) \in G_i$.

DEFINITION 3.1. If $\phi \in M$, we define $\mu(\phi)$ by

$$\begin{aligned} \mu(\phi) = \inf\{\mu > 0 &: \phi_i(x) \text{ can be expressed as in (3.3),} \\ &\text{and } |r| \leqq \mu \text{ for all } (r, \sigma) \in G_i \text{ and for } 1 \leqq i \leqq n\}. \end{aligned} \tag{3.4}$$

If $\phi \in M_+$, we define $\nu(\phi)$ by

$$\begin{aligned} \nu(\phi) = \inf\{\nu > 0 &: \phi_i(x) \text{ can be expressed as in (3.3),} \\ &\text{and } 0 \leqq r \leqq \nu \text{ for all } (r, \sigma) \in G_i \text{ and for } 1 \leqq i \leqq n\}. \end{aligned} \tag{3.5}$$

If C is a cone in a Banach space X and A and B are bounded linear maps of X to X, we will say that $A \leqq B$ if $(B - A)(C) \subset C$; the ordering depends on C. If $A(C) \subset C$ and $B(C) \subset C$, we will say that A and B are comparable if there exist positive numbers c_1 and c_2 such that

$$c_1 A \leqq B \leqq c_2 A.$$

If $C = K$ and $X = \mathbb{R}^n$, then bounded linear maps are $n \times n$ matrices, $A = (a_{ij}) \leqq B = (b_{ij})$ if and only if $a_{ij} \leqq b_{ij}$ for all i, j. If $a_{ij} \geqq 0$ and $b_{ij} \geqq 0$ for all i, j, A and B are comparable if and only if there exist positive reals c_1 and c_2 such that

$$c_1 a_{ij} \leqq b_{ij} \leqq c_2 a_{ij} \quad \text{for all } i, j.$$

Our next lemma is easy, but we give a proof for completeness.

LEMMA 3.1. *Assume that C is a cone in a Banach space X and that $A, B : X \to X$ are bounded linear operators such that $A(C) \subset C$ and $B(C) \subset C$. Assume that A and B are comparable; then there exist positive reals c_1 and c_2 such that*

$$c_1 A \leqq B \leqq c_2 A. \tag{3.6}$$

If A has finite projective diameter $\Delta(A)$ (see (1.6)), then B has finite projective diameter and

$$\Delta(B) \leqq \Delta(A) + 2 \log (c_2/c_1). \tag{3.7}$$

Proof. If x and y are any two elements of C such that Bx and By are nonzero, then by using (3.6) we see that Ax and Ay are nonzero. By definition of finite projective diameter, it follows that there exist positive reals α and β such that

$$\alpha A(x) \leqq A(y) \leqq \beta A(x) \quad \text{and} \quad \log(\beta/\alpha) \leqq \Delta(A). \tag{3.8}$$

By using (3.6) repeatedly we obtain from (3.8) that

$$\alpha(c_1/c_2)B(x) \leqq B(y) \leqq \beta(c_2/c_1)B(x), \tag{3.9}$$

which implies that

$$d(Bx, By) \leqq 2\log(c_2/c_1) + \log(\beta/\alpha). \tag{3.10}$$

Formulae (3.8) and (3.10) yield (3.7). □

In [26] it is proved that if $f \in \mathcal{M}$ (see § 1 for definitions), then f is C^∞ on $\mathring{K}$ and $f'(x)$ and $f'(y)$ are comparable for all $x, y \in \mathring{K}$ (this is not hard). We need a more precise version of this fact, relating the sizes of $f'(x)$ and $f'(y)$, when $f \in M$.

LEMMA 3.2. *Let K denote the standard cone in $\mathbb{R}^n$* (see (3.1)), *let $v = (1, 1, \cdots, 1)$ be the vector all of whose components are* 1, *and define $\psi \in K^*$ by*

$$\psi(x) = \sum_{i=1}^{n} x_i.$$

Suppose that $f: \mathring{K} \to \mathring{K}$ is homogeneous of degree 1 *and order-preserving. If $x \in \mathring{K}$, $\psi(x) = n$, and $d(x, v) \leqq R$ (d denotes Hilbert's projective metric), then*

$$e^{-R} f(v) \leqq f(x) \leqq e^{R} f(v). \tag{3.11}$$

If $f \in M$ and $\mu(f) < \gamma$ (see Definition 3.1), *then for all $x \in \mathring{K}$ such that $d(x, v) \leqq R$,*

$$\exp(-R(\gamma+1)) f'(v) \leqq f'(x) \leqq \exp(R(\gamma+1)) f'(v), \tag{3.12}$$

where $f'(x)$ denotes the Jacobian matrix at x.

Proof. If $\psi(x) = n = \psi(v)$ and $\alpha = m(x/v)$ and $\beta = M(x/v)$, we must have $\alpha \leqq 1 \leqq \beta$; and if $d(x, v) \leqq R$, then

$$\beta/\alpha \leqq e^R. \tag{3.13}$$

Formula (3.13) implies that $\beta \leqq e^R$ and $\alpha \geqq e^{-R}$ (since $\alpha \leqq 1 \leqq \beta$), and (3.11) follows from the homogeneity and order-preserving properties of f.

If $f \in M$, $\mu(f) < \gamma$, and f_i denotes the ith component of $f(x)$, then we can write

$$f_i(x) = \sum_{(r,\sigma) \in G_i} c_{ir\sigma} M_{r\sigma}(x), \qquad |r| < \gamma \quad \text{for } (r, \sigma) \in G_i. \tag{3.14}$$

Here G_i is a finite collection of ordered pairs (r, σ) with $r \in \mathbb{R}$ and σ a probability vector and $c_{ir\sigma} > 0$ for $(r, \sigma) \in G_i$, $1 \leqq i \leqq n$. If $d(x, v) \leqq R$ we have

$$e^{-R} \leqq x_j/x_k \leqq e^R \quad \text{for all } j, k. \tag{3.15}$$

A calculation implies that for $d(x, v) \leqq R$,

$$\frac{\partial M_{r\sigma}}{\partial x_j}(x) = \sigma_j [x_j (M_{r\sigma}(x))^{-1}]^{r-1}. \tag{3.16}$$

Recall that $M_{r\sigma}$ is an order-preserving map on $\mathring{K}$ for any real number r and that (3.15) implies that

$$e^{-R} x_j v \leqq x \leqq e^R x_j v.$$

Using this we conclude that for $d(x, v) \leqq R$ we have

$$e^{-R} \leqq x_j(M_{r\sigma}(x))^{-1} \leq e^R;$$

(3.16) then implies that

$$\sigma_j e^{-R|r-1|} \leqq \frac{\partial M_{r\sigma}}{\partial x_j}(x) \leqq \sigma_j e^{R|r-1|}. \tag{3.17}$$

Because $|r| < \gamma$ for all r such that $(r, \sigma) \in G_i$, we conclude from (3.17) that

$$\exp(-R(\gamma+1)) \sum_{(r,\sigma)\in G_i} c_{ir\sigma}\sigma_j \leqq \frac{\partial f_i}{\partial x_j}(x) \leqq \exp(R(\gamma+1)) \sum_{(r,\sigma)\in G_i} c_{ir\sigma}\sigma_j. \tag{3.18}$$

Of course (3.18) is equivalent to

$$\exp(-R(\gamma+1)) \frac{\partial f_i}{\partial x_j}(v) \leqq \frac{\partial f_i}{\partial x_j}(x) \leqq \exp(R(\gamma+1)) \frac{\partial f_i}{\partial x_j}(v),$$

which implies (3.2). □

We can now state a weak ergodic theorem for functions $f_j \in M$. In the statement of the following theorem recall that a nonnegative $n \times n$ matrix B is called "primitive" if there exists $p \geqq 1$ such that B^p has all positive entries.

THEOREM 3.1. *Let K be the standard cone in $\mathbb{R}^n$ (see (3.1)) and suppose that for $j \geqq 1$, $f_j \in M$, where M is the class of maps of $\mathring{K}$ into itself defined in § 1. Assume that $\mu(f_j) < \gamma < \infty$ for all $j \geqq 1$ (see Definition 3.1). If $v = (1, 1, \cdots, 1)$, suppose also that there exist an $n \times n$ primitive matrix B and an $n \times n$ matrix A such that $B \leqq f_j'(v) \leqq A$ for all $j \geqq 1$. Finally, assume that $\langle f_j \rangle$ satisfies the bounded orbit property with respect to Hilbert's projective metric d (see Definition 2.1). Then if $F_k = f_k f_{k-1} \cdots f_1$,*

$$\lim_{k\to\infty} d(F_k(x), F_k(v)) = 0 \quad \textit{for all } x \in \mathring{K},$$

$$\lim_{k\to\infty} \| F_k(x)\|F_k(x)\|^{-1} - F_k(v)\|F_k(v)\|^{-1}\| = 0 \quad \textit{for all } x \in \mathring{K}.$$

Proof. Select an integer $p \geqq 1$ such that B^p has all positive entries, and for this p let G_k and g_k be as defined in Theorem 2.1. The bounded orbit property implies that $\{F_k(x): k \geqq 0\}$ has finite projective diameter for any $x \in \mathring{K}$. In particular, there exists $R > 0$ such that

$$\{F_k(v): k \geqq 0\} \subset B_R(v) = \{z: d(z, v) < R\}. \tag{3.19}$$

For notational convenience, define

$$v_k = F_k(v) \quad \text{and} \quad u_k = G_k(u)/\|G_k(u)\|.$$

Because each f_j is nonexpansive with respect to d we see that if $x \in B_R(v_k)$, then for $j \geqq 1$ we have

$$f_{k+j} f_{k+j-1} \cdots f_{k+1}(x) \in B_R(v_{k+j}) \subset B_{2R}(v). \tag{3.20}$$

If $x \in B_R(u_k) = B_R(G_k(u))$, then by using (3.20) and the chain rule we see that

$$g_{k+1}'(x) = f_{kp+p}'(y_1) f_{kp+p-1}'(y_2) \cdots f_{kp+1}'(y_p), \tag{3.21}$$

where $y_1, y_2, \cdots, y_p$ are points in $B_{2R}(v)$ that depend on x and the maps f_j.

We now use Lemma 3.2 and (3.21) to conclude that

$$\exp(-p(\gamma+1)(2R)) f_{jp}'(v) f_{jp-1}'(v) \cdots f_{jp-p+1}'(v) \leqq g_j'(x),$$

$$g_j'(x) \leqq \exp(p(\gamma+1)(2R)) f_{jp}'(v) f_{jp-1}'(v) \cdots f_{jp-p+1}'(v)$$

for all $x \in B_R(u_{j-1})$, $j \geqq 1$. It follows that for some $c_1 > 0$ we have

$$\exp(-2p(\gamma+1)R)B^p \leqq g_j'(x) \leqq c_1 \exp(2p(\gamma+1)R)B^p \quad \text{for all } x \in B_R(u_{j-1}). \tag{3.22}$$

By using (3.22) we see that there exists a positive real number c, independent of j, such that

$$g_j'(x) \geqq c g_j'(u_{j-1}) \quad \text{for all } j \geqq 1, \quad x \in B_R(u_{j-1}),$$

which is (2.25).

If we define $B_j = g_j'(u_{j-1})$, then (3.22) and Lemma 3.1 imply that

$$\Delta(B_j) \leqq 4p(\gamma+1)R + \Delta(B^p),$$

so

$$\sum_{j=1}^{\infty} \exp(-\Delta(B_j)) = \infty.$$

The conclusions of Theorem 3.1 now follow directly from Corollary 2.1. □

As an immediate consequence of Theorems 3.1 and 2.4 we obtain the following result.

COROLLARY 3.1. *Let the notation and assumptions be as in Theorem* 3.1. *In addition, assume that there exists* $u \in \mathring{K}$ *such that* (2.40) *is satisfied. Then for each* $x \in \mathring{K}$ *there exists* $\gamma = \gamma(x) > 0$ *such that*

$$\lim_{m\to\infty} \| F_m(x) - \gamma F_m(u) \| = 0,$$

and $x \to \gamma(x)$ *is continuous, order-preserving, and homogeneous of degree* 1.

Similarly, by using Lemmas 3.1 and 3.2, we can derive the following corollary of Theorem 2.3. Details are left to the reader.

COROLLARY 3.2. *Let notation and assumptions be as in Theorem* 3.1. *Assume that there exists* $w \in K$ *such that*

$$\lim_{j\to\infty} d(f_j(w), w) = 0.$$

Then it follows that for all $x \in \mathring{K}$,

$$\lim_{k\to\infty} d(F_k(x), w) = 0 \quad \textit{and} \quad \lim_{k\to\infty} \| F_k(x) \| F_k(x) \|^{-1} - w \| = 0.$$

If σ is a probability vector and $x \in \mathring{K}$, we will use the notation

$$x^{\sigma} = \prod_{j=1} x_j^{\sigma_j}, \tag{3.23}$$

where x_j and σ_j are the jth components of x and σ, respectively. If $x, y \in \mathring{K}$ we will also use the notation

$$\log(x) = (\log(x_1), \log(x_2), \cdots, \log(x_n)) \quad \text{and} \quad y \cdot \log(x) = \sum_{j=1}^{n} y_j \log(x_j).$$

If $\langle f_k \rangle$ is a sequence of maps, order-preserving and homogeneous of degree 1, of $\mathring{K}$ into itself, verifying the bounded orbit property may be difficult. However, we now show that if $f_{ki}(x)$ can be bounded below in a suitable way by cx^{σ}, where $c > 0$, σ is a probability vector, and both c and σ may depend on k and i, then we can prove that $\langle f_k \rangle$ satisfies the bounded orbit property with respect to d. This idea has already been used in § 4 of [23] for the case where $f_k = f$ for all $k \geqq 1$.

Our next lemma is a slight variant of Lemma 4.1 in [23]. Since the argument is the same, the proof is left to the reader.

LEMMA 3.3. *Suppose that K is the standard cone in $\mathbb{R}^n$ and f and g are maps of $\mathring{K}$ into itself. Suppose that A is a nonnegative $n \times n$ matrix with no zero rows; suppose also that if $a_{ij} > 0$, then there exist a positive constant c and a probability vector σ (σ depends on i and j) such that $\sigma_j \geqq \eta > 0$ (where σ_j denotes the jth component of σ) and*

$$f_i(x) = \text{the } i\text{th component of } f(x) \geqq cx^{\sigma} \quad \text{for all } x \in \mathring{K}.$$

The constants c and η are assumed independent of i and j. Similarly, suppose that B is a nonnegative $n \times n$ matrix with no zero rows and that if $b_{ij} > 0$, then there exist a positive constant d and a probability vector τ such that $\tau_j \geqq \theta > 0$ and

$$g_i(y) \geqq dy^{\tau} \quad \text{for all } y \in \mathring{K}.$$

Here d and θ are assumed independent of i and j, but τ may depend on i and j. Then BA is an $n \times n$ nonnegative matrix with no zero rows, and if the entry in row i, column j of BA is nonzero, there exists a probability vector μ such that

$$g_i(f(x)) \geqq \lambda x^{\mu} \quad \text{for all } x \in \mathring{K},$$

where $\lambda \geqq cd$ and $\mu_j \geqq \theta\eta$.

In the statement of the following theorem, recall that $f_{ki}(x)$ denotes the ith component of a map $f_k : \mathring{K} \to \mathring{K}$.

THEOREM 3.2. *Let K denote the standard cone in $\mathbb{R}^n$ and suppose that $f_k : \mathring{K} \to \mathring{K}$, $k \geqq 1$, is a sequence of maps that are order-preserving and homogeneous of degree 1. Let $A = (a_{ij})$ be an $n \times n$ nonnegative, primitive matrix. If $a_{ij} > 0$, assume that there exist $c > 0$ and $\delta > 0$ (independent of i, j, and $k \geqq 1$) and a probability vector σ with $\sigma_j \geqq \delta$ (σ may depend on i, j, and k) such that*

$$f_{ki}(x) \geqq cx^{\sigma} \quad \text{for all } x \in \mathring{K}.$$

If $v = (1, 1, \cdots, 1)$, assume that there exists a constant c_2 such that $f_k(v) \leqq c_2 v$ for all $k \geqq 1$. Then $\langle f_k \rangle$ satisfies the bounded orbit property with respect to Hilbert's projective metric d, and for every $u \in \mathring{K}$, there exists $R > 0$ such that

$$d(F_k(u), u) \leqq R \quad \text{for all } k \geqq 1.$$

Proof. Let F_k be as defined in Theorem 2.1 and select $p \geqq 1$ such that all entries of A^p are positive. For any fixed $k \geqq 0$, define $g : \mathring{K} \to \mathring{K}$ by

$$g = f_{k+p} f_{k+p-1} \cdots f_{k+1}. \tag{3.24}$$

By applying Lemma 3.3 $(p-1)$ times and recalling that A^p has all positive entries, we see that for any i, j with $1 \leqq i, j \leqq n$ there exists a probability vector σ (depending on i, j, k, and p) with $\sigma_j \geqq \delta^p = \eta$ (σ_j = the jth component of σ) and

$$g_i(x) = \text{the } i\text{th component of } g(x) \geqq c^p x^{\sigma} \quad \text{for all } x \in \mathring{K}. \tag{3.25}$$

Suppose that $x \in \mathring{K}$ and $d(x, v) \leqq R$, and select j so that

$$x_j = M(x/v) = M. \tag{3.26}$$

If $m = m(x/v)$, we obtain from (3.25) and (3.26) that

$$g_i(x) \geqq c^p x^{\sigma} \geqq c^p M^{\sigma_j} m^{1-\sigma_j} \geqq c^p (M/m)^{\eta} m = c^p M^{\eta} m^{1-\eta}. \tag{3.27}$$

In deriving (3.27) we use $M/m \geqq 1$.

On the other hand, we assume that $f_j(v) \leqq c_2 v$ for all j, so

$$g(v) \leqq c_2^p v. \tag{3.28}$$

Because $x \leqq Mv$ we conclude from (3.28) that

$$g_s(x) \leqq c_2^p M \quad \text{for } 1 \leqq s \leqq n. \tag{3.29}$$

Combining (3.27) and (3.29) we conclude that

$$d(g(x), v) \leqq p \log\left(\frac{c_2}{c}\right) + (1-\eta)\log\left(\frac{M}{m}\right) \leqq p \log\left(\frac{c_2}{c_1}\right) + (1-\eta)R. \tag{3.30}$$

It follows from (3.30) that there exist a real number λ, $0 < \lambda < 1$, and a number R_1, both independent of k in the definition of g, such that

$$d(g(x), v) \leqq \lambda R \quad \text{if } d(x, v) \leqq R \quad \text{and} \quad R \geqq R_1. \tag{3.31}$$

We can also assume that R_1 is so large that

$$d(F_j(v), v) < R_1 \quad \text{for } 1 \leqq j < p. \tag{3.32}$$

In general we can write F_m for $m \geqq p$ in the form

$$F_m = g_1 g_2 \cdots g_t F_j, \tag{3.33}$$

where $0 \leqq j < p$ and each g_i in (3.33) is assumed to be of the form given by (3.24) for some $k \geqq 0$. Since every g as in (3.24) maps $B_{R_1}(v)$ into itself and since $F_j(v) \in B_{R_1}(v)$ for $0 \leqq j < p$, we conclude that $F_m(v) \in B_{R_1}(v)$ for all $m \geqq 1$. □

With the aid of Theorems 3.1 and 3.2 we can give a more concrete weak ergodic theorem.

THEOREM 3.3. *Let K denote the standard cone in $\mathbb{R}^n$ (see (3.1)) and let $\langle f_k \rangle$, $k \geqq 1$, be a sequence of maps such that $f_k \in M_+$ for all $k \geqq 1$. Assume that $\mu(f_k) < \rho < \infty$ for all $k \geqq 1$ (see Definition 3.1). Let $v = (1, 1, \cdots, 1)$ denote the vector all of whose components equal 1 and assume that there exists an $n \times n$ nonnegative, primitive matrix B such that*

$$f_k'(v) \geqq B \quad \text{for all } k \geqq 1. \tag{3.34}$$

Assume that there exists $\beta_2 > 0$ such that

$$f_k(v) \leqq \beta_2 v \quad \text{for all } k \geqq 1. \tag{3.35}$$

Then we have that for all $x \in \mathring{K}$

$$\lim_{k\to\infty} d(F_k(x), F_k(v)) = 0,$$

$$\lim_{k\to\infty} \|F_k(x)\|F_k(x)\|^{-1} - F_k(v)\|F_k(v)\|^{-1}\| = 0.$$

If there exists $u \in \mathring{K}$ such that

$$\sup\{\|F_k(u)\| : k \geqq 1\} < \infty \quad \text{and} \quad \inf\{\|F_k(u)\| : k \geqq 1\} > 0,$$

then for every $x \in \mathring{K}$ there exists $\gamma = \gamma(x) > 0$ such that

$$\lim_{k\to\infty} \|F_k(x) - \gamma F_k(u)\| = 0.$$

If there exists $w \in \mathring{K}$ such that

$$\lim_{j\to\infty} d(f_j(w), w) = 0,$$

then for every $x \in \mathring{K}$,

$$\lim_{k\to\infty} d(F_k(x), w) = 0.$$

Proof. The existence of a matrix A as in Theorem 3.1 follows from (3.35) and the equation $f_k'(v)(v)=f_k(v)$. Thus, by virtue of Theorem 3.1 and Corollaries 3.1 and 3.2, it suffices to prove that the sequence $\langle f_k\rangle$ satisfies the bounded orbit property. We establish the bounded orbit property by using Theorem 3.2.

First note that, because f_k is homogeneous of degree 1,

$$f_k(v)=f_k'(v)(v)\geqq Bv;$$

therefore there exist positive constants β_1 and β_2 so that

$$\beta_1 v\leqq f_k(v)\leqq \beta_2 v \quad \text{for all } k\geqq 1. \tag{3.36}$$

If we write

$$f_{ki}(x)=\sum_{(r,\sigma)\in\Gamma_{ki}} c_{kir\sigma}M_{r\sigma}(x), \quad 1\leqq i\leqq n, \quad k\geqq 1,$$

we can assume $r\geqq 0$ for $(r,\sigma)\in\Gamma_{ki}$ (because $f_k\in M_+$). Formula (3.36) implies that

$$\beta_1\leqq f_{ki}(v)=c_{ki}=\sum_{(r,\sigma)\in\Gamma_{ki}} c_{kir\sigma}\leqq\beta_2. \tag{3.37}$$

It is a classical result (see [18]) that $M_{r\sigma}(x)\geqq M_{0\sigma}(x)$ for $r\geqq 0$, so

$$f_{ki}(v)\geqq c_{ki}\sum_{(r,\sigma)\in\Gamma_{ki}} c_{kir\sigma}c_{ki}^{-1}M_{0\sigma}(x). \tag{3.38}$$

If we apply log to both sides of (3.38) and use the concavity of log we obtain

$$\log f_{ki}(x)\geqq(\log c_{ki})+\left(\sum_{(r,\sigma)\in\Gamma_{ki}} c_{kir\sigma}c_{ki}^{-1}\sigma\right)\cdot(\log x). \tag{3.39}$$

If we define a probability vector τ_{ki} by

$$\tau_{ki}=\sum_{(r,\sigma)\in\Gamma_{ki}} c_{kir\sigma}c_{ki}^{-1}\sigma, \tag{3.40}$$

we obtain from (3.37), (3.39), and (3.40) that

$$f_{ki}(x)\geqq c_{ki}x^{\tau_{ki}}\geqq\beta_1 x^{\tau_{ki}}. \tag{3.41}$$

Denote by b_i the ith row of the matrix B. A simple calculation shows that the ith row of the Jacobian matrix $f_k'(v)$ is $c_{ki}\tau_{ki}$, so by the hypotheses of our theorem we have that

$$\tau_{ki}\geqq c_{ki}^{-1}b_i\geqq\beta_2^{-1}b_i. \tag{3.42}$$

If $B=(b_{ij})$, define a positive number δ by

$$\delta=\inf\{b_{ij}\beta_2^{-1}\colon b_{ij}>0\}.$$

Then it follows from (3.41) and (3.42) that if $b_{ij}>0$, the jth component of τ_{ki} in (3.41) is greater than or equal to δ. Since β_1 and δ are independent of i, j and $k\geqq 1$, Theorem 3.2 implies that $\langle f_k\rangle$ satisfies the bounded orbit property. □

Remark 3.1. Note that, for functions f_k that satisfy the hypotheses of Theorem 3.3, it can easily happen that f_k does not map certain nonzero points in the boundary of K into the interior of K and that the diameter (with respect to Hilbert's projective metric d) of $\{f_k(x)\colon x\in\mathring{K}\}$ is *not* finite. Both of these phenomena are illustrated by the simple arithmetic-geometric mean map

$$f_k(x)=f(x)=\left(\frac{x_1+x_2}{2},\ \sqrt{x_1x_2}\right).$$

In addition, both phenomena are typical of many examples of interest. For example, suppose that $x_k = F_k(x)$, where F_k is as in Theorem 3.3. More generally, for $f_k \in M_+$, we can define $\tilde{f}_k$ by

$$\tilde{f}_k(x) = \mu_k(x) f_k(x),$$

where $\mu_k(x)$ is a positive scalar function of x. If we define $\tilde{F}_k = \tilde{f}_k \tilde{f}_{k-1} \cdots \tilde{f}_1$ and $x_k = \tilde{F}_k(x)$ and $x_0 = x$, it is easy to see that

$$\tilde{F}_k(x) = \lambda_k(x) F_k(x), \qquad \lambda_k(x) = \prod_{j=1}^{k} \mu_j(x_{j-1}).$$

Here $\lambda_k(x)$ is a positive scalar, and the presence of $\lambda_k(x)$ does not affect the validity of the first part of Theorem 3.3, because Hilbert's projective metric does not distinguish points on rays. In this terminology, x may represent an initial "population vector" (so that the j-component of x represents the number of members of the population in class j) and $x_k = \tilde{F}_k(x)$ may represent the population vector at time k. Under reasonable assumptions on the biological model, we expect f_k to vanish at certain nonzero points on the boundary of K. Note, however, that if f_k is linear and irreducible, f_k does not vanish on nonzero points of the boundary of K. This point indicates a drawback of linear weak ergodic theorems in population biology.

REFERENCES

[1] J. ARAZY, T. CLAESSON, S. JANSON, AND J. PEETRE, *Means and their iterations*, in Proc. Nineteenth Nordic Congress of Mathematicians, Reykjavik, 1984, Icelandic Mathematical Society, 1985, pp. 191-212.

[2] F. L. BAUER, *An elementary proof of the Hopf inequality for positive operators*, Numer. Math., 7 (1965), pp. 331-337.

[3] H. BAUER AND H. S. BEAR, *The part metric in convex sets*, Pacific J. Math., 30 (1969), pp. 15-33.

[4] G. BIRKHOFF, *Extensions of Jentzsch's theorems*, Trans. Amer. Math. Soc., 85 (1957), pp. 219-227.

[5] ———, *Uniformly semi-primitive multiplicative processes*, Trans. Amer. Math. Soc., 104 (1962), pp. 37-51.

[6] P. BUSHELL, *Hilbert's metric and positive contraction mappings in Banach space*, Arch. Rational Mech. Anal., 52 (1973), pp. 330-338.

[7] ———, *On the projective contraction ratio for positive linear mappings*, J. London Math. Soc., 6 (1973), pp. 256-258.

[8] ———, *On a class of Volterra and Fredholm nonlinear integral equations*, Math. Proc. Cambridge Philos. Soc., 79 (1976), pp. 329-335.

[9] ———, *The Cayley-Hilbert metric and positive operators*, Linear Algebra Appl., 84 (1986), pp. 271-280.

[10] H. CASWELL AND D. E. WEEKS, *Two sex models: chaos, extinction and other dynamic consequences of sex*, Amer. Naturalist, 128 (1986), pp. 707-735.

[11] J. E. COHEN, *Ergodic theorems in demography*, Bull. Amer. Math. Soc., 1 (1979), pp. 275-295.

[12] ———, *Random arithmetic-geometric mean and random pi: observations and conjectures*, preprint.

[13] J. E. COHEN AND R. D. NUSSBAUM, *Arithmetic-geometric means of positive matrices*, Math. Proc. Cambridge Philos. Soc., 101 (1987), pp. 209-219.

[14] C. M. DAFERMOS AND M. SLEMROD, *Asymptotic behavior of non-linear contraction semigroups*, J. Funct. Anal., 13 (1973), pp. 97-106.

[15] C. J. EVERETT AND N. METROPOLIS, *A generalization of the Gauss limit for iterated means*, Adv. in Math., 7 (1971), pp. 297-300.

[16] T. FUJIMOTO AND U. KRAUSE, *Asymptotic properties for inhomogeneous iterations of nonlinear operators*, SIAM J. Math. Anal., 19 (1988), pp. 841-853.

[17] M. GOLUBITSKY, E. KEELER, AND M. ROTHSCHILD, *Convergence of the age structure: application of the projective metric*, Theoret. Population Biol., 7 (1975), pp. 84-93.

[18] G. H. HARDY, J. E. LITTLEWOOD, AND G. POLYA, *Inequalities*, Cambridge University Press, Cambridge, 1934.

[19] E. HOPF, *An inequality for positive linear integral operators*, J. Math. Mech., 12 (1963), pp. 683-692.

[20] M. A. KRASNOSELSKII AND A. V. SOBOLEV, *Spectral clearance of a focusing operator*, Funct. Anal. Appl., 17 (1983), pp. 58-59.

[21] U. KRAUSE, *Perron's stability theorem for non-linear mappings*, J. Math. Econom., 15 (1986), pp. 275-282.

[22] K. C. LAND AND A. ROGERS, EDS., *Multidimensional Mathematical Demography*, Academic Press, New York, 1982.

[23] R. D. NUSSBAUM, *Convexity and* log *convexity for the spectral radius*, Linear Algebra Appl., 73 (1986), pp. 59-122.

[24] ———, *Iterated nonlinear maps and Hilbert's projective metric: a summary*, in Dynamics of Infinite Dimensional Sytems, S.-N. Chow and J. K. Hale, eds., NATO-ASI Series F, Vol. 37, Springer-Verlag, Berlin, New York, 1987, pp. 231-248.

[25] ———, *Iterated nonlinear maps and Hilbert's projective metric*, Mem. Amer. Math. Soc., Vol. 75, No. 391 (1988).

[26] ———, *Iterated nonlinear maps and Hilbert's projective metric*, II, Mem. Amer. Math. Soc., Vol. 79, No. 401 (1989).

[27] A. M. OSTROWSKI, *Positive matrices and functional analysis*, in Recent Advances in Matrix Theory, H. Schneider, ed., University of Wisconsin Press, Madison, WI, 1964, pp. 81-101.

[28] R. A. POLLAK, *Demography's two-sex problem*, preprint.

[29] A. J. B. POTTER, *Applications of Hilbert's projective metric to certain classes of non-homogeneous operators*, Quart. J. Math., 28 (1977), pp. 93-99.

[30] ———, *Hilbert's projective metric applied to a class of positive operators*, in Ordinary and Partial Differential Equations, Lecture Notes in Math. 564, Springer-Verlag, Berlin, New York, pp. 377-382.

[31] P. A. SAMUELSON, *Generalizing Fisher's "reproductive value": nonlinear, homogeneous, biparental systems*, Proc. Nat. Acad. Sci. U.S.A., 74 (1977), pp. 5772-5775.

[32] H. SCHAEFER, *Topological Vector Spaces*, Springer-Verlag, Berlin, New York, 1971.

[33] ———, *Banach Lattices and Positive Operators*, Springer-Verlag, Berlin, New York, 1974.

[34] R. SCHOEN, *The two-sex multi-ethnic stable population model*, Theoret. Population Biol., 129 (1986), pp. 343-364.

[35] E. SENETA, *Nonnegative Matrices and Markov Chains*, Second edition, Springer-Verlag, Berlin, New York, 1981.

[36] A. C. THOMPSON, *On certain contraction mappings in a partially ordered vector space*, Proc. Amer. Math. Soc., 14 (1963), pp. 438-443.

[37] P. P. ZABREIKO, M. A. KRASNOSELSKII, AND YU. V. POKORNYI, *On a class of positive linear operators*, Funktsional. Anal. i Prilozhen., 5 (1971), pp. 9-17; Funct. Anal. Appl. (1972), pp. 272-279.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 461-482, March 1990
© 1990 Society for Industrial and Applied Mathematics
010

ON BERNSTEIN-SZEGÖ ORTHOGONAL POLYNOMIALS ON SEVERAL INTERVALS*

FRANZ PEHERSTORFER†

Abstract. Let $l\in\mathbb{N}$, $a_1<a_2<\cdots<a_{2l}$, $E_l=\bigcup_{k=1}^{l}[a_{2k-1},a_{2k}]$, $H(x)=\prod_{k=1}^{2l}(x-a_k)$ and let $\rho_\nu(x)=c\prod_{k=1}^{\nu^*}(x-w_k)^{\nu_k}$ be a real polynomial with $w_k\notin\operatorname{int}(E_l)$ for $k=1,\cdots,\nu^*$ and $\nu_k=1$ if w_k is a boundary point of E_l. For given ρ_ν and $\varepsilon=(\varepsilon_1,\cdots,\varepsilon_{\nu^*})$, $\varepsilon_k\in\{-1,1\}$, the following linear functional on $\mathbb{P}$, $\mathbb{P}$ denoting the space of real polynomials, is defined:

$$\Psi_{H,\rho_\nu,\varepsilon}(p)=\int_{E_l}p(x)\frac{\sqrt{-H(x)}}{\rho_\nu(x)}\operatorname{sgn}\left(-\prod_{k=1}^{l}(x-a_{2k-1})\right)dx$$

$$+\sum_{k=1}^{\nu^*}(1-\varepsilon_k)\sum_{j=1}^{\nu_k}\mu_{j,k}p^{(j-1)}(w_k)$$

where $\mu_{j,k}$'s are certain numbers. Polynomials orthogonal with respect to the (not necessarily positive definite) linear functional $\Psi_{H,\rho_\nu,\varepsilon}$ are characterized. Those polynomials are given the name Bernstein-Szegö orthogonal polynomials on several intervals. Special attention is given to the most interesting case $\varepsilon=(1,\cdots,1)$.

Key words. orthogonal polynomial, associated polynomial, several intervals, Bernstein-Szegö weight function, point measure

AMS(MOS) subject classifications. 42C05, 33A65

1. Introduction and notation.

Let $l\in\mathbb{N}$, $a_k\in\mathbb{R}$ for $k=1,\cdots,2l$, $a_1<a_2<\cdots<a_{2l}$ and put

$$E_l=\bigcup_{k=1}^{l}[a_{2k-1},a_{2k}],\qquad H(x)=\prod_{k=1}^{2l}(x-a_k),$$

$$\frac{1}{h(x)}=\begin{cases}\operatorname{sgn}\left(\prod_{k=1}^{l}(x-a_{2k-1})\right)\Big/\sqrt{-H(x)} & \text{for } x\in E_l,\\ 0 \quad \text{elsewhere.}\end{cases}$$

R and S denote polynomials with leading coefficient 1 of degree r, respectively, s, that satisfy the relation

$$R(x)S(x)=H(x)$$

and ρ_ν denotes a real polynomial of degree ν that has no zero in E_l, i.e.,

$$\rho_\nu(x)=c\prod_{k=1}^{\nu^*}(x-w_k)^{\nu_k}$$

where $c\in\mathbb{R}\setminus\{0\}$, $\nu^*\in\mathbb{N}_0$, $\nu_k\in\mathbb{N}$ for $k=1,\cdots,\nu^*$, $\nu=\sum_{k=1}^{\nu^*}\nu_k$, $w_k\in\mathbb{C}\setminus E_l$, for $k=1,\cdots,\nu^*$, and the w_k's are real or appear in pairs of complex numbers. Furthermore, we set

$$\rho_{\nu,k}(x)=\rho_\nu(x)/(x-w_k)^{\nu_k}\quad\text{for } k=1,\cdots,\nu^*,$$

$$\Xi_{\nu^*}=\{(\varepsilon_1,\cdots,\varepsilon_{\nu^*}):\varepsilon_k\in\{-1,+1\}\quad\text{for } k=1,\cdots,\nu^*\}.$$

* Received by the editors August 24, 1987; accepted for publication (in revised form) November 18, 1988.

† Institut für Mathematik, J. Kepler Universität Linz, A-4040 Linz, Austria.

For given R, ρ_ν, and $\varepsilon \in \Xi_{\nu^*}$ we define the following linear functionals on the space of real polynomials $\mathbb{P}$:

$$L_{R,\rho_\nu,\varepsilon}(p) = \sum_{k=1}^{\nu^*} \frac{1-\varepsilon_k}{(\nu_k-1)!} \left(\frac{pR}{\rho_{\nu,k}\sqrt{H}}\right)^{(\nu_k-1)}(w_k) \quad \text{for } p \in \mathbb{P}, \tag{1.1}$$

$$\Psi_{R,\rho_\nu,\varepsilon}(p) = \int p \frac{R}{\rho_\nu h}\, dx + L_{R,\rho_\nu,\varepsilon}(p) \quad \text{for } p \in \mathbb{P} \tag{1.2}$$

where we make the additional assumption that $\varepsilon_{k+1} = \varepsilon_k$ if w_k and w_{k+1} are complex conjugate and where that branch of $\sqrt{H}$ is chosen (see Lemma 1) that is analytic on $\mathbb{C}\setminus E_l$ and satisfies

$$\operatorname{sgn}\sqrt{H(y)} = \operatorname{sgn} \prod_{k=1}^{l} (y - a_{2k-1}) \quad \text{for } y \in \mathbb{R}\setminus E_l.$$

Here $g^{(j)}$ denotes the jth derivative of g. Instead of $\Psi_{1,\rho,\varepsilon}$ we write $\Psi_{\rho,\varepsilon}$.

In this paper we study polynomials $p_n = x^n + \cdots$, $n \in \mathbb{N}$, which are orthogonal with respect to $\Psi_{R,\rho_\nu,\varepsilon}$, i.e., which satisfy

$$\Psi_{R,\rho_\nu,\varepsilon}(x^j p_n) = 0 \quad \text{for } j = 0, \cdots, \tilde{n}$$

where $\tilde{n} \geqq n-1$. Let us note that $\tilde{n} > n-1$ is possible since the linear functional need not be definite. It is not difficult to demonstrate (see Lemma 2) that for given R, ρ_ν, ε there exists a unique sequence of so-called basic integers (i_n), $i_0 := 0 < i_1 < i_2 < \cdots$, with $i_{n+1} \leqq i_n + \nu + l$ and a unique sequence of polynomials $p_{i_n} = x^{i_n} + \cdots$, $n \in \mathbb{N}_0$, such that

$$\Psi_{R,\rho_\nu,\varepsilon}(x^j p_{i_n}) = 0 \quad \text{for } j = 0, \cdots, i_{n+1}-2, \tag{1.3}$$

$$\Psi_{R,\rho_\nu,\varepsilon}(x^{i_{n+1}-1} p_{i_n}) \neq 0, \tag{1.4}$$

and the p_{i_n}'s satisfy a recurrence relation of the type

$$p_{i_n}(x) = d_{i_n}(x) p_{i_{n-1}}(x) - \lambda_{i_n} p_{i_{n-2}}(x) \quad \text{for } n \in \mathbb{N} \tag{1.5}$$

where $\lambda_{i_n} \in \mathbb{R}\setminus\{0\}$, $d_{i_n} \in \mathbb{P}_{i_n - i_{n-1}}$ (as usual, $\mathbb{P}_n$ denotes the set of polynomials of degree at most n), $p_{i_0} = 1$, and $p_{i_{-1}} = 0$. If the functional $\Psi_{R,\rho_\nu,\varepsilon}$ can be represented in the form

$$\Psi_{R,\rho_\nu,\varepsilon}(p) = \int p\, d\psi_{R,\rho_\nu,\varepsilon} \tag{1.6}$$

where $\psi_{R,\rho_\nu,\varepsilon}$ is a (not necessarily positive) measure, we write also, instead of (1.3), that $p_{i_n} \perp \mathbb{P}_{i_{n+1}-2}$ with respect to $d\psi_{R,\rho_\nu,\varepsilon}$, respectively, with respect to $R/\rho_\nu h$, if $L_{R,\rho_\nu,\varepsilon} = 0$, i.e., if $\varepsilon_k = 1$ for $k = 1, \cdots, \nu^*$. Obviously, (1.6) holds if all zeros of ρ_ν are simple and real. The orthogonality condition in this case becomes

$$\int_{E_l} \frac{x^j p_{i_n}(x)}{\rho_\nu(x)} \sqrt{\frac{\prod_{\mu\in J}|x-a_\mu|}{\prod_{\mu\in K\setminus J}|x-a_\mu|}} \operatorname{sgn}\left(\prod_{\mu\in I}(x-a_\mu)\right) \operatorname{sgn}\left(\prod_{\mu\in J}(x-a_\mu)\right) dx$$

$$+ \sum_{k=1}^{\nu^*} \frac{(1-\varepsilon_k)}{\rho_\nu'(w_k)} \sqrt{\frac{\prod_{\mu\in J}(w_k-a_\mu)}{\prod_{\mu\in K\setminus J}(w_k-a_\mu)}}\, w_k^j p_{i_n}(w_k) = 0 \quad \text{for } j = 0, \cdots, i_{n+1}-2 \tag{1.7}$$

where $K = \{1, 2, \cdots, 2l\}$, $I = \{1, 3, 5, \cdots, 2l-1\}$ and $J \subset K$. If $\psi_{R,\rho_\nu,\varepsilon}$ from (1.6) is a positive measure, i.e., if $\Psi_{R,\rho_\nu,\varepsilon}$ is positive definite, then it follows from (1.6) and (1.3) that the sequence of basic integers (i_n) satisfies $i_n = n$ for $n \in \mathbb{N}$, which implies that the recurrence relation (1.5) becomes

$$p_n(x) = (x-\alpha_n) p_{n-1}(x) - \lambda_n p_{n-2}(x) \quad \text{for } n \in \mathbb{N}. \tag{1.8}$$

For the case where (p_n) satisfies a recurrence relation of the form (1.8) with $\lambda_n \neq 0$ for $n \in \mathbb{N}$ (i.e., for the case where $i_n = n$ for $n \in \mathbb{N}$), we define the so-called associated polynomials $(p_n^{(j)})_{n \in \mathbb{N}_0}$ of order $j, j \in \mathbb{N}_0$, by

$$p_n^{(j)}(x) = (x - \alpha_{n+j}) p_{n-1}^{(j)}(x) - \lambda_{n+j} p_{n-2}^{(j)} \quad \text{for } n \in \mathbb{N} \tag{1.9}$$

with $p_0^{(j)} = 1$ and $p_{-1}^{(j)} = 0$.

To state our results we need the following additional notation. For $p \in \mathbb{P}$ let

$$p_{R,\rho_\nu,\varepsilon}^{[1]}(z) = \Psi_{R,\rho_\nu,\varepsilon}\left(\frac{p(z) - p(x)}{z - x}\right) \tag{1.10}$$

where it is understood that $\Psi_{R,\rho_\nu,\varepsilon}$ operates on x. The index R,ρ_ν,ε is omitted if there is no possibility of confusion. Note that $p_n^{[1]}$, $n \in \mathbb{N}$, is, up to a constant, the associated polynomial of degree $n-1$ of order 1 if $\Psi_{R,\rho_\nu,\varepsilon}$ is definite and if p_n denotes the corresponding orthogonal polynomial with leading coefficient 1. ∂p denotes the exact degree of $p \in \mathbb{P}$. Henceforth p_n, p_{i_n}, q_m, q_{i_m} denote polynomials of degree n, i_n, m, i_m, respectively with leading coefficient 1.

For the single interval case, say $[-1, +1]$, $\Psi_{R,\rho_\nu,\varepsilon}$ with $\varepsilon = (1, 1, \cdots, 1)$ is of the type

$$\int_{-1}^{+1} . (1-x)^\alpha (1+x)^\beta / \rho_\nu(x)\, dx \quad \text{where } \alpha, \beta \in \{-\tfrac{1}{2}, \tfrac{1}{2}\}. \tag{1.11}$$

Polynomials orthogonal with respect to such weight functions have been studied by Bernstein [5] and Szegö [25, pp. 31-33] and are now known under those authors' names. These polynomials also play an important role in L^p- and Chebyshev approximation on $[-1, +1]$ (see [1, pp. 249-254] and [6]). Since $\Psi_{R,\rho_\nu,\varepsilon}$ can be considered a generalization of (1.11), we call polynomials orthogonal with respect to $\Psi_{R,\rho_\nu,\varepsilon}$ Bernstein-Szegö polynomials on several intervals. Another extension of the Bernstein-Szegö orthogonal polynomials for a more general class of weight functions than that given in (1.11) has been given by Nevai [26].

Polynomials orthogonal on two intervals with respect to R/h, their recursion coefficients, and their close connection with polynomials deviating least from zero on two intervals with respect to the maximum, respectively, the L^1-, norm are investigated by Peherstorfer in [21] and [22]. For the representation of L^1-minimal polynomials on several intervals in terms of polynomials orthogonal with respect to R/h, see [20]. Using elliptic, respectively, Abelian, functions, polynomials orthogonal with respect to particular weight functions of the type (1.7), without point measure, have been studied by Achieser and Tomčuk [2], [3], Nuttall and Singh [18, § 5], [19, § 4.3.1-§4.3.4], and Magnus [17, § 4].

In [13] Geronimus (see also [11]) has shown that polynomials satisfying a three-term recurrence relation with periodic recursion coefficients are orthogonal on a set of several intervals with respect to a functional of the type $\Psi_{H,\rho,\varepsilon}$. Polynomials orthogonal to certain special weight functions of the type above have been given by Ismail [15], very recently by Geronimo and Van Assche [12, Remark 7], and by Peherstorfer [23]. In this paper we do not use elliptic, respectively, Abelian, functions, and our approach, and our results as well, are different from those given in the above-mentioned papers.

This paper is organized as follows. In § 2 we list and discuss the main results of the paper. Section 3 contains all the preliminary results. In the remaining sections the listed results of § 2 are proved.

Finally, we would like to mention that, based on the characterization of the orthogonal polynomials in § 2, we will show in a forthcoming paper how to obtain a

nonlinear recurrence relation for the recursion coefficients of the polynomials orthogonal with respect to $\Psi_{R,\rho,\varepsilon}$, and we will give a necessary and sufficient condition for the periodicity of the recursion coefficients of orthogonal polynomials.

2. Statement of results. The first main result is Theorem 1.

THEOREM 1. *Let $n \in \mathbb{N}$ be such that $n \geqq \max\{\nu+1-l, (\nu+1-r)/2\}$, put $m = n+r-l$, and suppose that $p_n \in \mathbb{P}_n$ satisfies*

$$Rp_n^2 = Sq_m^2 + \rho_\nu \tag{2.1}$$

where $q_m \in \mathbb{P}_m$. Furthermore, assume that p_n and q_m have no common zero and that at the zeros w_k of ρ_ν

$$Rp_n(w_k) = \varepsilon_k \sqrt{H(w_k)}\, q_m(w_k) \quad \text{for } k = 1, \cdots, \nu^*$$

where $\varepsilon_k \in \{-1, +1\}$, and set $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_{\nu^*})$. Then the following propositions hold:

$$\text{(a)} \quad \Psi_{R,\rho_\nu,\varepsilon}(x^j p_n) = 0 \text{ for } j = 0, \cdots, n+l-2;$$

$$\text{(b)} \quad \Psi_{S,\rho_\nu,\varepsilon}(x^j q_m) = 0 \text{ for } j = 0, \cdots, m+l-2; \tag{2.2}$$

$$\text{(c)} \quad q_m(z) = \int \frac{(Rp_n)(z) - (Rp_n)(x)}{z-x} \frac{dx}{h(x)};$$

and, if $m > \nu - l$,

$$p_n(z) = \int \frac{(Sq_m)(z) - (Sq_m)(x)}{z-x} \frac{dx}{h(x)}, \tag{2.3}$$

$$\text{(d)} \quad q_m(z) = \begin{cases} \displaystyle\int \frac{p_n(z)-p_n(x)}{z-x} \frac{R(x)}{h(x)}\, dx & \text{if } r < l, \\ \displaystyle\int \frac{p_n(z)-p_n(x)}{z-x} \frac{R(x)}{h(x)}\, dx + p_n(z) & \text{if } r = l. \end{cases}$$

COROLLARY 1. *Suppose that the assumptions of Theorem 1 are fulfilled and that $\rho_\nu = \tilde{\rho}_{\nu-\mu} g_\mu$, where $\mu \in \{0, \cdots, \nu\}$. Furthermore, assume that w_{k_j}, $j = 1, \cdots, \nu^* - \mu^*$, are the zeros of $\tilde{\rho}_{\nu-\mu}$ and put $\tilde{\varepsilon} = (\varepsilon_{k_1}, \cdots, \varepsilon_{k_{\nu^*-\mu^*}})$. Then*

$$\Psi_{R,\tilde{\rho}_{\nu-\mu},\tilde{\varepsilon}}(x^j p_n) = 0 \quad \text{for } j = 0, \cdots, n+l-2-\mu,$$

$$\Psi_{S,\tilde{\rho}_{\nu-\mu},\tilde{\varepsilon}}(x^j q_m) = 0 \quad \text{for } j = 0, \cdots, m+l-2-\mu.$$

In particular, $p_n \perp \mathbb{P}_{n+l-2-\nu}$ ($q_m \perp \mathbb{P}_{m+l-2-\nu}$) with respect to R/h (S/h).

The next corollary shows how to construct a polynomial ρ for a given polynomial p_n such that p_n is orthogonal with respect to $\Psi_{\rho,\varepsilon}$.

COROLLARY 2. *Let $p_n \in \mathbb{P}_n$. Put*

$$q_m(z) = \int \frac{p_n(z)-p_n(x)}{z-x} \frac{dx}{h(x)} \quad \text{and} \quad \rho = p_n^2 - Hq_m^2.$$

Furthermore, suppose that p_n and q_m have no common zero and that at the zeros w_k of ρ

$$p_n(w_k) = \varepsilon_k \sqrt{H(w_k)}\, q_m(w_k).$$

Then $\rho \in \mathbb{P}_{n+l-1}$ and

$$\Psi_{\rho,\varepsilon}(x^j p_n) = 0 \quad \text{for } j = 0, \cdots, n+l-2.$$

In view of Theorem 1 the question arises: under which conditions does $\Psi_{R,\rho_\nu,\varepsilon}(p) = \int p(R/\rho_\nu h)\, dx$ for $p \in \mathbb{P}$? We give the answer for the special case $R = 1$ only, since the conditions become rather complicated in the general case.

THEOREM 2. *Let $n \in \mathbb{N}$, $n \geqq \max\{l, \nu\}$ and suppose that $p_n \in \mathbb{P}_n$ satisfies*

$$p_n^2 - Hq_m^2 = \rho_\nu \tag{2.1}$$

where p_n and q_m, $q_m \in \mathbb{P}_m$, have no common zero. $p_n \perp \mathbb{P}_{n+l-2}$ with respect to $1/\rho_\nu h$ if and only if all zeros of p_n and q_m are simple and are in $\operatorname{int}(E_l)$ and the zeros of p_n and $\prod_{k=2}^{l}(x-a_{2k-1})q_m$ interlace strictly.

The next theorem shows that the sufficiency conditions of Theorem 1 are also necessary. For Theorem 3 compare also [17, p. 161].

THEOREM 3. *Let i_n be a basic integer with respect to $\Psi_{R,\rho_\nu,\varepsilon}$ and suppose that $i_{n+1} + i_n + r \geqq \nu + l + 1$. Furthermore, let $v \in \mathbb{P}_{\nu-1}$ and $u \in \mathbb{P}_{r-(l+\nu)}$ be such that at the zeros w_k of $\rho_\nu(x) = \prod_{k=1}^{\nu^*}(x-w_k)^{\nu_k}$*

$$v^{(j)}(w_k) = \varepsilon_k (R/\sqrt{H})^{(j)}(w_k) \quad \text{for } j = 0, \cdots, \nu_k - 1, \tag{2.4}$$

$$R(z)/\rho_\nu(z)\sqrt{H(z)} = u(z) + O(z^{-1})$$

and put

$$Y = u\rho_\nu + v.$$

Then the following propositions hold:

(a) $$Rp_{i_n}^2 - S(Yp_{i_n} + \rho_\nu p_{i_n}^{[1]})^2 = \rho_\nu g_{(i_n)}$$

where $g_{(i_n)} \in \mathbb{P}_{i_n+l-i_{n+1}}$ and $g_{(i_n)}$ has leading coefficient $2\Psi_{R,\rho_\nu,\varepsilon}(x^{i_{n+1}-1}p_{i_n})$.

(b) *Y is of the form*

$$Y(z) = \int \frac{R(z)-R(x)}{z-x}\frac{dx}{h(x)} - \Psi_{R,\rho_\nu,\varepsilon}\left(\frac{\rho_\nu(z)-\rho_\nu(x)}{z-x}\right).$$

(c) *If $i_{n+1} \geqq \nu + 1$, then*

$$\Psi_{S,\rho_\nu,\varepsilon}(x^j(Yp_{i_n} + \rho_\nu p_{i_n}^{[1]})) = 0 \quad \text{for } j = 0, \cdots, i_{n+1} + r - l - 2.$$

Notation. Let (i_n) be a sequence of basic integers with respect to $\Psi_{R,\rho_\nu,\varepsilon}$ and let Y be defined as in Theorem 3. We put

$$q_{i_m} = Yp_{i_n} + \rho_\nu p_{i_n}^{[1]} \tag{2.5}$$

where $i_m = i_n + r - l$. Note that, in view of Theorem 3 and Theorem 1, (i_m) is a sequence of basic integers with respect to $\Psi_{S,\rho_\nu,\varepsilon}$ for those m satisfying $i_{m+1} \geqq \nu + 1 + r - l$ and that the q_{i_m}'s satisfy the same recurrence relation as the p_{i_n}'s.

As an easy consequence of Theorem 3 we obtain that a polynomial orthogonal with respect to $\Psi_{R,\rho,\varepsilon}$ satisfies a second-order differential equation. Using different methods, this fact has been proved by several authors [4], [10], [14].

COROLLARY 3. *Suppose that the assumptions of Theorem 3 are fulfilled and let p_{i_n}, ρ_ν, $g_{(i_n)}$, Y be such as in Theorem 3. Put $N = R\rho_\nu g_{(i_n)}$ and $\mu = \partial g_{(i_n)}$. Assume that Rp_{i_n} and q_{i_m} have no common zero.*

(a) *There is an $u \in \mathbb{P}_{l+\nu+\mu-1}$ such that $Ruq_{i_m} = -N'Rp_{i_n}/2 + N(Rp_{i_n})'$.*

(b) *$y = Rp_{i_n}/\sqrt{N}$ satisfies the differential equation $2G(x)y'' + G(x)y' + 2y = 0$, where $G = -H(N/Ru)^2$.*

Note that the differential equation of Corollary 3(b) is not of Sturm–Liouville type, since G depends on p_{i_n}.

The next theorem and corollary show how to make those polynomials orthogonal with respect to $\Psi_{R,\rho_\nu\tilde{\rho}_\mu,\sigma}$ if we know all polynomials $(p_{i_n})_{n\in\mathbb{N}}$ and one polynomial $\tilde{p}_j$ orthogonal with respect to $\Psi_{R,\rho_\nu,\varepsilon}$, respectively, $\Psi_{R,\tilde{\rho}_\mu,\tilde{\varepsilon}}$, where $\sigma = (\varepsilon_1, \cdots, \varepsilon_{\nu^*}, \pm\tilde{\varepsilon}_1, \cdots, \pm\tilde{\varepsilon}_{\mu^*})$.

THEOREM 4. *Let (i_n) be the sequence of basic integers with respect to $\Psi_{R,\rho_\nu,\varepsilon}$ and suppose that $i_{n+1}+i_n+r \geqq \nu+l+1$. Assume that*

$$u_j^2 - Hv_k^2 = \tilde{\rho}_\mu \tag{2.6}$$

where the polynomials $u_j = x^j + \cdots \in \mathbb{P}_j$ and $v_k = x^k + \cdots \in \mathbb{P}_k$ have no common zero, $2j > \mu$, and at the zeros $\tilde{w}_\kappa$ of $\tilde{\rho}_\mu$

$$u_j(\tilde{w}_\kappa) = \tilde{\varepsilon}_\kappa \sqrt{H(\tilde{w}_\kappa)}\, v_k(\tilde{w}_\kappa) \quad \text{for } \kappa = 1, \cdots, \mu^*,$$

$\tilde{\varepsilon}_\kappa \in \{-1, +1\}$. Furthermore, put

$$P_{i_n+j} = u_j p_{i_n} + S v_k q_{i_m} \quad \text{and} \quad Q_{i_m+j} = u_j q_{i_m} + R v_k p_{i_n}, \tag{2.7}$$

$$\tilde{P}_{i_n+\mu-j} = u_j p_{i_n} - S v_k q_{i_m} \quad \text{and} \quad \tilde{Q}_{i_m+\mu-j} = u_j q_{i_m} - R v_k p_{i_n} \tag{2.8}$$

where $i_m = i_n + r - l$. Then the following propositions hold:

(a) $$\Psi_{R,\rho_\nu\tilde{\rho}_\mu,\sigma}(x^\lambda P_{i_n+j}) = 0 \quad \text{for } \lambda = 0, \cdots, i_{n+1}+j-2$$

$$\Psi_{S,\rho_\nu\tilde{\rho}_\mu,\sigma}(x^\lambda Q_{i_m+j}) = 0 \quad \text{for } \lambda = 0, \cdots, i_{m+1}+j-2$$

where $\sigma = (\varepsilon_1, \cdots, \varepsilon_{\nu^*}, \tilde{\varepsilon}_1, \cdots, \tilde{\varepsilon}_{\mu^*})$.

(b) *If $2i_n + \mu + r > 2j + \nu$ then $\tilde{P}_{i_n+\mu-j}$, respectively, $\tilde{Q}_{i_m+\mu-j}$, is a polynomial of degree $i_n+\mu-j$, respectively, $i_m+\mu-j$, with leading coefficient $\tilde{K}$, where $\tilde{K}$ is the leading coefficient of $\tilde{\rho}_\mu$, that satisfies for $i_{n+1} > \nu+j$*

$$\Psi_{R,\rho_\nu\tilde{\rho}_\mu,\sigma}(x^\lambda \tilde{P}_{i_n+\mu-j}) = 0 \quad \text{for } \lambda = 0, \cdots, i_{n+1}+\mu-j-2,$$

respectively,

$$\Psi_{S,\rho_\nu\tilde{\rho}_\mu,\tilde{\sigma}}(x^\lambda \tilde{Q}_{i_m+\mu-j}) = 0 \quad \text{for } \lambda = 0, \cdots, i_{m+1}+\mu-j-2$$

where $\tilde{\sigma} = (\varepsilon_1, \cdots, \varepsilon_{\nu^*}, -\tilde{\varepsilon}_1, \cdots, -\tilde{\varepsilon}_{\mu^*})$.

(c) *The polynomials P_{i_n+j}, $\tilde{P}_{i_n+\mu-j}$, Q_{i_m+j}, and $\tilde{Q}_{i_m+\mu-j}$ satisfy the same recurrence relation with respect to i_n as the polynomials p_{i_n} for $i_{n+1}+i_n+r \geqq \nu+l+1$.*

Since in general not $\tilde{\rho}_\mu$, but $\tilde{\rho}_\mu g$, $g \in \mathbb{P}_{l-1}$, has a representation of the form (2.6), let us state the following corollary.

COROLLARY 4. *Suppose that the assumptions of Theorem 4 are fulfilled. In addition, assume that $i_n = n$ for $n \in \mathbb{N}_0$ and that $\tilde{\rho}_\mu = \bar{\rho}_{\mu-\tau} g$, where $\bar{\rho}_{\mu-\tau}$, $g \in \mathbb{P}$ and $\partial g = \tau$. Then for each $n \geqq \nu+\mu-j$ there are τ numbers $c_{0,n}, \cdots, c_{\tau,n}$, such that*

$$\bar{P}_{n+j} := \sum_{\kappa=0}^{\tau} c_{\kappa,n} P_{n+j+\kappa}/g \quad \text{and} \quad \bar{Q}_{m+j} := \sum_{\kappa=0}^{\tau} c_{\kappa,n} Q_{m+j+\kappa}/g$$

are polynomials of degree at most $n+j$, respectively, $m+j$, that satisfy

$$\Psi_{R,\rho_\nu\bar{\rho}_{\mu-\tau},\delta}(x^\lambda \bar{P}_{n+j}) = 0 \quad \text{for } \lambda = 0, \cdots, n+j-1,$$

$$\Psi_{S,\rho_\nu\bar{\rho}_{\mu-\tau},\delta}(x^\lambda \bar{Q}_{m+j}) = 0 \quad \text{for } \lambda = 0, \cdots, m+j-1$$

where $\delta = (\varepsilon_1, \cdots, \varepsilon_{\nu^*}, \bar{\varepsilon}_1, \cdots, \bar{\varepsilon}_{(\mu-\tau)^*})$.

Concerning the single interval case $[-1, +1]$, we obtain from Theorem 4 the corollary below.

COROLLARY 5. *Let* $E_1=[-1,1]$ *and let* $t_\nu(z)=\prod_{\kappa=1}^{\nu^*}(z-z_\kappa)^{\nu_\kappa}$ *and* $\tilde{t}_\mu(z)=\prod_{\kappa=1}^{\mu^*}(z-z_\kappa)^{\mu_\kappa}$, *where* $|z_\kappa|<1$, *respectively,* $|\tilde{z}_\kappa|>1$ *for* $\kappa=1,\cdots,\nu^*$, *respectively,* $\kappa=1,\cdots,\mu^*$, *and put*

$$\sum_{\kappa=0}^{\nu+\mu} d_\kappa z^\kappa = t_\nu(z)\tilde{t}_\mu(z),$$

$$\rho_\nu(x)=|t_\nu(e^{i\phi})|^2 \quad \textit{and} \quad \tilde{\rho}_\mu(x)=|\tilde{t}_\mu(e^{i\phi})|^2$$

$x=\cos\phi$, $\phi\in[0,\pi]$. *Then for* $n\geqq\nu+\mu$

$$\Psi_{\rho_\nu\tilde{\rho}_\mu,\sigma}\left(x^\lambda\sum_{\kappa=0}^{\nu+\mu} d_\kappa T_{n-(\nu+\mu)+\kappa}\right)=0 \quad \textit{for } \lambda=0,\cdots,n-1$$

and for $n>\nu+\mu$

$$\Psi_{x^2-1,\rho_\nu\tilde{\rho}_\mu,\sigma}\left(x^\lambda\sum_{\kappa=0}^{\nu+\mu} d_\kappa U_{n-1-(\nu+\mu)+\kappa}\right)=0 \quad \text{for } \lambda=0,\cdots,n-2$$

where $\sigma=(\sigma_1,\cdots,\sigma_{\nu^*+\mu^*})$ *is such that*

$$\sigma_\kappa=\begin{cases}1 & \textit{at the zeros of } \rho_\nu,\\ -1 & \textit{at the zeros of } \tilde{\rho}_\mu.\end{cases}$$

As usual the polynomials T_n, *respectively,* U_n, $n\in\mathbb{N}$, *denote the Chebyshev polynomials of degree n of first, respectively, second, kind on* $[-1,+1]$.

Remark. In Corollary 5 we have excluded the case that the polynomials t_ν, respectively, $\tilde{t}_\mu$, have zeros on the circumference $|z|=1$. The reason is the following simple fact. If $\sum_{\kappa=0}^{\nu} d_\kappa z^\kappa$, $d_\kappa\in\mathbb{R}$, has the complex conjugate zeros $e^{i\tilde{\phi}}$, $e^{-i\tilde{\phi}}$ then $\sum_{\kappa=0}^{\nu} d_\kappa T_{n-\nu+\kappa}(\sum_{\kappa=0}^{\nu} d_\kappa U_{n-1-\nu+\kappa})$ and $\rho_\nu(x)=|\sum_{\kappa=0}^{\nu} d_\kappa e^{i\kappa\phi}|^2$ have the common zero $\tilde{x}=\cos\tilde{\phi}$.

Let us note that, by a well-known result of Fejér and Riesz, each polynomial ρ_ν that is positive on $[-1,+1]$ can be represented uniquely in the form ($x=\cos\phi$, $\phi\in(0,\pi)$)

$$\rho_\nu(x)=c\left|\prod_{\kappa=1}^{\nu}(z-z_\kappa)\right|^2=\tilde{c}\left|\prod_{\kappa=1}^{\nu}(1-zz_\kappa)\right|^2$$

where c, $\tilde{c}\in\mathbb{R}^+$, and $|z_\kappa|<1$ for $\kappa=1,\cdots,\nu$. Note that the second equality gives the unique representation of ρ_ν by a polynomial of degree at most ν with all its zeros outside of the open unit disk.

Applying Corollary 5 to $z-b$, respectively, $(z-b-\sqrt{b^2+1})(z-b+\sqrt{b^2+1})$, $b\in\mathbb{R}$, we get the orthogonality properties of U_n-bU_{n-1}, respectively, $U_n-2bU_{n-1}-U_{n-2}$, given by Chihara [8], [7, p. 205] (see also [29]). Note that there is a misprint in [7, form. (13.6)]. The denominator $1+b^2-t$ in (13.6) should be replaced by $1+b^2-t^2$. From a different point of view, the single interval case has also been considered by Geronimus [13] and by Dombrowski and Nevai [28].

Next let us determine that functional to which the associated polynomials are orthogonal.

THEOREM 5. *Let* $(p_{i_n})_{n\in\mathbb{N}_0}$ *be orthogonal with respect to* $\Psi_{R,\rho_\nu,\varepsilon}$ *and suppose that the basic integers* (i_n) *satisfy* $i_n=n$ *for each* $n\in\mathbb{N}_0$. *Let* $j\in\mathbb{N}$, $2j-2>\nu+l-r$, $k=j+r-l$, *and let*

$$Rp_{j-1}^2-Sq_{k-1}^2=\rho_\nu g_{(j-1)} \tag{2.9}$$

where at the zeros $z_{\kappa,j-1}$, $\kappa=1,\cdots,l^*-1$, *of* $g_{(j-1)}$

$$(Rp_{j-1})(z_{\kappa,j-1})=\delta_{\kappa,j-1}(\sqrt{H}q_{k-1})(z_{\kappa,j-1}),\qquad \kappa=1,\cdots,l^*-1.$$

Furthermore, we put for $n\in\mathbb{N}_0$

$$C_{n+l}^{(j)}=(Rp_{n+j}p_{j-1}-Sq_{n+k}q_{k-1})/K_j\rho_\nu,$$

$$D_n^{(j)}=(q_{n+k}p_{j-1}-p_{n+j}q_{k-1})/K_j\rho_\nu$$

where $K_j=\prod_{\mu=1}^{j}\lambda_\mu$ *and* $\lambda_1=\Psi_{R,\rho_\nu,\varepsilon}(1)$.

(a) $D_n^{(j)}=p_n^{(j)}$ *for* $n\in\mathbb{N}_0$.

(b) $C_{n+l}^{(j)}$, $n\in\mathbb{N}_0$, *are polynomials of degree* $n+l$ *with leading coefficient* 1 *that satisfy the recurrence relation*

$$C_{n+l}^{(j)}(x)=(x-\alpha_{n+j})C_{n+l-1}^{(j)}(x)-\lambda_{n+j}C_{n+l-2}^{(j)}(x)\quad \text{for } n\geqq 2.$$

(c) *The polynomials* $C_{n+l}^{(j)}$, *respectively,* $p_n^{(j)}$, $n\in\mathbb{N}_0$, *are orthogonal with respect to* $\Psi_{g_{(j-1)},-\delta_{j-1}}$, *respectively,* $\Psi_{H,g_{(j-1)},-\delta_{j-1}}$.

Remark. If $\Psi_{R,\rho_\nu,\varepsilon}$ is positive definite, then (if we use the facts that $R/\rho_\nu h>0$ by (1.2) and $R\rho_\nu g_{(j-1)}>0$ on int (E_l) by (2.9)), $g_{(j-1)}$ has exactly one zero in each interval $[a_{2\kappa}, a_{2\kappa+1}]$, $\kappa=1,\cdots,l-1$.

Finally, we describe the location of the zeros of p_n and q_m if $\Psi_{\rho,\varepsilon}$ can be represented by a positive absolutely continuous measure.

THEOREM 6. *Assume that* $1/h\rho_\nu>0$ *on* E_l *and let* $\rho_\nu(x)=\prod_{k=1}^{\mu}(x-w_k)v(x)$, *where* $a_2<w_1<w_2<\cdots<w_\mu<a_{2l-1}$ *and* v *is a polynomial that is positive on* $[a_1, a_{2l}]$. *Let* ρ_1, ρ_2 *be polynomials such that* $\rho_1\rho_2=\prod_{k=1}^{\mu}(x-w_k)$ *and that* ρ_1, *respectively,* ρ_2, *vanishes at the first, third, etc., respectively second, fourth, etc., zero of* ρ_ν *in* (a_{2j}, a_{2j+1}), $j=1,\cdots,l-1$. *Furthermore, suppose that* $n\geqq\max\{\nu+1-l,\ (\nu+1-r)/2\}$ *and that,* $m=n+r-l$,

$$p_n^2-Hq_m^2=\rho_\nu g_{(n)} \tag{2.10}$$

where $g_{(n)}\in\mathbb{P}_{l-1}$ *and* p_n *and* q_m *have no common zero. Then* $p_n\perp\mathbb{P}_{n-1}$ ($q_m\perp\mathbb{P}_{m-1}$) *with respect to* $1/\rho_\nu h$ (h/ρ_ν) *if and only if all zeros of* p_n *and* q_m *are in* (a_1, a_{2l}), p_n *and* q_m *have at most one zero in each interval* (a_{2j}, a_{2j+1}), $j=1,\cdots,l-1$, *and the zeros of* $\rho_1 q_m$ *and* $\rho_2 p_n$ *interlace strictly.*

3. Preliminary results. To prove our results we need several lemmas. Lemma 1 plays an important role.

LEMMA 1. *Let* $t\in\mathbb{P}_l$ *with leading coefficient* 1.

(a) *For* $z\in\mathbb{C}\setminus E_l$

$$\int\frac{t(x)}{z-x}\frac{dx}{h(x)}=\begin{cases}\dfrac{t(z)}{\sqrt{H(z)}} & \text{if } \partial t\leqq l-1,\\[2ex] \dfrac{t(z)}{\sqrt{H(z)}}-1 & \text{if } \partial t=l\end{cases}$$

where that branch of $\sqrt{\ }$ *is chosen for which*

$$\lim_{z\to\infty}\frac{z^l}{\sqrt{H(z)}}=1.$$

(b) *If* $y \in \mathbb{R}\setminus E_l$ *and* $\partial t \leqq l-1$, *then*

$$\int \frac{t(x)}{y-x}\frac{dx}{h(x)} = \frac{t(t)}{\sqrt{H(y)}} = \frac{t(y)\operatorname{sgn}\prod_{k=1}^{l}(y-a_{2k-1})}{\sqrt{|H(y)|}}.$$

(c) $$\int x^j \frac{dx}{h(x)} = \begin{cases} 0 & \text{for } j = 0, \cdots, l-2, \\ 1 & \text{for } j = l-1. \end{cases}$$

Proof. (a) First let us consider the case $\partial t \leqq l-1$. Put

$$C = \bigcup_{k=1}^{2l-1} (a_k, a_{k+1}),$$

$$F(z) = t(z)/\sqrt{H(z)} = t(z)/\sqrt{|H(z)|}\, e^{i(\arg H(z))/2} \text{ for } z \in \mathbb{C}\setminus E_l,$$

$$F^+(x) = \lim_{z\to x \text{ on } C^+} F(z) \quad \text{and} \quad F^-(x) = \lim_{z\to x \text{ on } C^-} F(z)$$

where $z \to x$ on $C^{\pm}$ means that $x \in C$, $|z-x| \to 0$ and $\operatorname{Im} z \gtrless 0$. From Theorem 4.1 and p. 495 of [16] it follows that

$$F(z) = \frac{1}{2\pi i}\int_C \frac{F^-(x)-F^+(x)}{z-x}\,dx \quad \text{for } z \in \mathbb{C}\setminus[a_1, a_{2l}].$$

Since by simple calculation

$$\lim_{z\to x \text{ on } C^{\pm}} \arg H(z) = \pm(2l-k)\pi \quad \text{for } x \in (a_k, a_{k+1}), \quad k \in \{0, \cdots, 2l\}$$

where $a_0 := -\infty$ and $a_{2l+1} := \infty$, we obtain

$$F^-(x)-F^+(x) = \frac{t(x)}{\sqrt{|H(x)|}}\left[\lim_{z\to x \text{ on } C^-} e^{-i(\arg H(z))/2} - \lim_{z\to x \text{ on } C^+} e^{-i(\arg H(z))/2}\right]$$

$$= \frac{t(x)}{\sqrt{|H(x)|}} \begin{cases} 2i(-1)^{l-k} & \text{for } x \in (a_{2k-1}, a_{2k}), \quad k \in \{1, \cdots, l\}, \\ 0 & \text{for } x \in (a_{2k}, a_{2k+1}), \quad k \in \{1, \cdots, l-1\}, \end{cases}$$

which proves (a) for $\partial t \leqq l-1$ and $z \in \mathbb{C}\setminus[a_1, a_{2l}]$. Observing that both functions

$$\int \frac{t(x)}{z-x}\frac{dx}{h(x)} \quad \text{and} \quad t(z)/\sqrt{H(z)}$$

are analytic on $\mathbb{C}\setminus E_l$, it follows by the Identity Theorem that equality holds for $z \in \mathbb{C}\setminus E_l$.

If $\partial t = l$ we get, setting

$$F_1(z) = F(z) - 1,$$

that

$$\lim_{z\to\infty} F_1(z) = 0.$$

If we proceed as above the assertion follows.

(b) We obtain by calculation that for $y \in (a_{2k}, a_{2k+1})$, $k \in \{0, \cdots, l\}$,

$$\lim_{z\to y} t(z)/\sqrt{H(z)} = (-1)^{l-k} t(y)/\sqrt{|H(y)|},$$

which, in conjunction with (a), gives the assertion.

Part (c) follows immediately by series expansion of $1/\sqrt{H(z)}$ at $z = \infty$.

Next let us state some facts on polynomials orthogonal with respect to a linear functional. The following two lemmas are essentially known (see [27] and [30]).

LEMMA 2. *Let $K \in \mathbb{N}_0$ and let c be a real linear functional on $\mathbb{P}$ with*

$$c(x^j) = 0 \quad \text{for } j = 0, \cdots, K-1 \text{ and } c(x^K) \neq 0.$$

Assume that there does not exist $t \in \mathbb{P}$ such that $c(x^j t) = 0$ for all $j \in \mathbb{N}_0$. Then the following propositions hold:

(a) *There exists a unique sequence of so-called basic integers i_n, $n \in \mathbb{N}_0$, $i_0 := 0 < i_1 < i_2 < \cdots$ and a sequence of polynomials $p_{i_n}(x) = x^{i_n} + \cdots$ such that for $n \in \mathbb{N}_0$*

$$c(x^j p_{i_n}) = 0 \quad \text{for } j = 0, \cdots, i_{n+1} - 2 \text{ and } c(x^{i_{n+1}-1} p_{i_n}) \neq 0.$$

(b) *The polynomials p_{i_n} satisfy a recurrence relation of the form*

$$p_{i_{n+1}}(x) = d_{i_{n+1}}(x) p_{i_n}(x) - \lambda_{i_{n+1}} p_{i_n}(x) \quad \text{for } n \in \mathbb{N} \tag{3.1}$$

where $d_{i_{n+1}} \in \mathbb{P}_{i_{n+1}-i_n}$, $\lambda_{i_{n+1}} \in \mathbb{R} \setminus \{0\}$ and $p_0(x) = 1$.

(c) *The associated polynomials*

$$p_{i_n}^{[1]}(z) = c\left(\frac{p_{i_n}(z) - p_{i_n}(x)}{z - x}\right)$$

satisfy the same recurrence relation as the p_{i_n}'s. Note that $p_{i_n}^{[1]} \in \mathbb{P}_{i_n - 1 - K}$.

(d) $p_{i_n} p_{i_{n+1}}^{[1]} - p_{i_{n+1}} p_{i_n}^{[1]} = c(x^{i_{n+1}-1} p_{i_n})$ *for* $n \in \mathbb{N}$.

Proof. Part (a) follows from [27, Props. 1.14, 1.17].

(b) Representing $p_{i_{n+1}}$ in the form $p_{i_{n+1}} = \sum_{j=0}^{n} e_{i_{j+1}} p_{i_j}$, where $e_{i_{j+1}} \in \mathbb{P}_{i_{j+1}-i_j}$, and using the orthogonality property of the polynomials p_{i_j} gives the assertion.

Part (c) follows with the help of (3.1).

(d) From (3.1) we get that for $n \in \mathbb{N}$

$$\lambda_{i_{n+1}} = c(x^{i_{n+1}-1} p_{i_n}) / c(x^{i_n - 1} p_{i_{n-1}}). \tag{3.2}$$

Since, again by the recurrence relation of p_{i_n} and $p_{i_n}^{[1]}$,

$$p_{i_n} p_{i_{n+1}}^{[1]} - p_{i_{n+1}} p_{i_n}^{[1]} = \lambda_{i_{n+1}} (p_{i_{n-1}} p_{i_n}^{[1]} - p_{i_n} p_{i_{n-1}}^{[1]}),$$

the assertion follows with the help of (3.2) by induction arguments.

LEMMA 3. *Suppose that the assumptions of Lemma 2 are fulfilled, put $c_j = c(x^j)$ for $j \in \mathbb{N}_0$, and assume that $\sum_{j=0}^{\infty} c_j z^{-(j+1)}$ converges for sufficiently large $|z|$. Furthermore, let $n \geqq K+1$, $p_n(z) = z^n + \cdots \in \mathbb{P}_n$, $q \in \mathbb{P}_{n-1-K}$, and $\mu \in \mathbb{N}$. Then*

$$\frac{q(z)}{p_n(z)} = \sum_{j=0}^{n+\mu-2} c_j z^{-(j+1)} + O(z^{-(n+\mu)}) \tag{3.3}$$

if and only if

$$c(x^j p_n) = 0 \quad \text{for } j = 0, 1, \cdots, \mu - 2,$$

$$q(z) = c\left(\frac{p_n(z) - p_n(x)}{z - x}\right).$$

Proof. Sufficiency. Since

$$\begin{aligned} \frac{c((p_n(z) - p_n(x))/(z-x))}{p_n(z)} &= c\left(\frac{1}{z-x}\right) - \frac{c(p_n/(z-x))}{p_n(z)} \\ &= c\left(\frac{1}{z-x}\right) - \frac{1}{z^\mu p_n(z)} c\left(\frac{x^{\mu-1} p_n}{1 - (x/z)}\right) \end{aligned}$$

where in the last equality we have used the orthogonality property of p_n, the sufficiency part is proved.

Necessity. Since p_n has leading coefficient 1 and $c_K \neq 0$, it follows that $\partial q = n-1-K$. If we set

$$p_n(x) = \beta_0 x^n + \beta_1 x^{n-1} + \cdots + \beta_n,$$

it follows immediately from (3.3) that

$$\sum_{k=0}^{n} c_{n+j-k}\beta_k = 0 \quad \text{for } j = 0, \cdots, \mu-2,$$

which is obviously equivalent to

$$c(x^j p_n) = 0 \quad \text{for } j = 0, \cdots, \mu-2.$$

Since, by the sufficiency part,

$$c\left(\frac{p_n(z)-p_n(x)}{z-x}\right) - q(z) = p_n(z)O(z^{-(n+\mu)}) = O(z^{-\mu}),$$

the lemma is proved.

LEMMA 4. *There exists an* $M \in \mathbb{R}^+$ *such that* (1.1) *holds for* $p := 1/(z-x)$, $z \in \{z \in \mathbb{C}: |z| > M\}$, *where* $L_{R,\rho_\nu,\varepsilon}$ *operates on* x.

Proof. Since $L_{R,\rho_\nu,\varepsilon}$ is of the form $\kappa := \max_{1 \leqq k \leqq \nu^*} \nu_k$,

$$L_{R,\rho_\nu,\varepsilon}(p) = \sum_{j=0}^{\kappa} \sum_{k=1}^{\nu^*} \mu_{j,k} p^{(j)}(w_k),$$

we get, setting $M_1 = \max_{j,k} |\mu_{j,k}|$ and $M = \max\{1, \max_k |w_k|\}$, that for $j > \kappa$

$$|L_{R,\rho_\nu,\varepsilon}(x^j)| \leqq j(j-1)\cdots(j-\kappa)M^j M_1 \nu^*(\kappa+1).$$

Thus

$$L_{R,\rho_\nu,\varepsilon}\left(\frac{1}{z-x}\right) = \sum_{j=0}^{\infty} L_{R,\rho_\nu,\varepsilon}(x^j) z^{-j-1} \quad \text{uniformly for } |z| > M. \tag{3.4}$$

Using Leibniz's formula we get by (1.1) that the series is equal to

$$\sum_{k=1}^{\nu^*} \frac{1-\varepsilon_k}{(\nu_k-1)!} \sum_{\mu=0}^{\nu_k-1} \binom{\nu_k-1}{\mu} \left(\frac{R}{\rho_{\nu,k}\sqrt{H}}\right)^{(\nu_k-1-\mu)}(w_k) \sum_{j=0}^{\infty} \frac{(x^j)^{(\mu)}(w_k)}{z^{j+1}}.$$

Taking into consideration the fact that for $|w_k/z| < 1$

$$\left(\frac{1}{z-x}\right)^{(\mu)}(w_k) = \sum_{j=0}^{\infty} \frac{(x^j)^{(\mu)}(w_k)}{z^{j+1}}$$

the assertion is proved.

DEFINITION 1 (see [24]). We say that a function f has n sign changes on a set X if there are $n+1$ points $\{x_i | i = 1, \cdots, n+1, x_i < x_{i+1}\}$ in X so that $f(x_i)(-1)^i > 0 (< 0)$. $S(f, X)$ denotes the number of sign changes on X.

LEMMA 5. *Let* $m \in \mathbb{N}_0$, $v \in L^1(E_l)$ *and let* $p_n = x^n + \cdots \in \mathbb{P}$. *If* $S(v, E_l) = m$ *and* $p_n \perp \mathbb{P}_{n-1+j}$, $j \in \mathbb{N}_0$, *on* E_l *with respect to* v, *then* $j \leqq m$ *and* p_n *has at least* $n+j-m$ *simple zeros in* $\operatorname{int}(E_l)$.

Proof. First we demonstrate that $j \leqq m$. Assuming that $j > m$ and choosing a $t \in \mathbb{P}_m$ such that

$$t \operatorname{sgn} v \geqq 0 \quad \text{on } E_l,$$

we get

$$\int_{E_l} p_n^2 tv\,dx > 0,$$

which is a contradiction to the orthogonality property of p_n.

Now suppose that p_n has $k < n+j-m$ simple zeros in int (E_l). Then p_n is of the form

$$p_n(x) = f_k(x)g(x)$$

where f_k, $g \in \mathbb{P}$ and $g \geqq 0$ on E_l. Since

$$\int_{E_l} x^i f_k g v\,dx = \int_{E_l} x^i p_n v\,dx = 0 \quad \text{for } i = 0, \cdots, n-1+j,$$

we get

$$f_k \perp \mathbb{P}_{k-1+(n+j-k)} \quad \text{on } E_l \text{ with respect to } gv.$$

Using the fact that

$$S(gv, E_l) = S(v, E_l) = m,$$

in view of the first part we obtain that

$$n+j-k \leqq m,$$

which is a contradiction to the assumption.

Remark. Observing that

$$\Psi_{R,\rho_\nu,\varepsilon}(\rho_\nu p) = \int_{E_l} \frac{pR}{h}\,dx \quad \text{for } p \in \mathbb{P},$$

with the help of Lemma 5 we get that two basic integers i_n, i_{n+1}, $n \in \mathbb{N}_0$, with respect to $\Psi_{R,\rho_\nu,\varepsilon}$, satisfy $i_{n+1} \leqq i_n + \nu + l$. Moreover, taking into account (3.4), Lemmas 2 and 3 can be applied to $\Psi_{R,\rho_\nu,\varepsilon}$.

4. Proofs of Theorems 1 and 3 and Corollaries 1–3. In this section we omit, because of abbreviation, the indices of p_n, p_{i_n}, q_m, $\cdots$, ρ_ν, respectively.

Proof of Theorem 1. (c) from (2.1) we get that

$$\frac{q(z)}{Rp(z)} = \frac{1}{\sqrt{H(z)}} + O(z^{-(2n+r+l-\nu)}), \tag{4.1}$$

which, in conjunction with Lemmas 3 and 1(a), gives

$$Rp \perp \mathbb{P}_{n+l-2-\nu} \quad \text{with respect to } 1/h \tag{4.2}$$

and that q is of the form (2.2).

From (4.1) it follows that

$$\frac{(Sq)(z)}{p(z)} = \sqrt{H(z)} + O(z^{-(2n+r+l-\nu)+2l}),$$

which implies, in view of $n+l = m+s$, that

$$\frac{p(z)}{Sq(z)} = \frac{1}{\sqrt{H(z)}} + O(z^{-(2m+s+l-\nu)}). \tag{4.3}$$

Hence

$$Sq \perp \mathbb{P}_{m+l-2-\nu} \quad \text{with respect to } 1/h, \tag{4.4}$$

and p is of the form (2.3).

(a) Since each $t \in \mathbb{P}_{n+l-2}$ can be represented in the form $t = u\rho + v$, where $u \in \mathbb{P}_{n+l-2-\nu}$ and $v \in \mathbb{P}_{\nu-1}$, it follows from (4.2) and the fact that $L_{R,\rho,\varepsilon}(u\rho p) = 0$ that

$$\Psi_{R,\rho,\varepsilon}(tp) = \Psi_{R,\rho,\varepsilon}(vp) = \int \frac{v}{\rho} pR \frac{dx}{h} + L_{R,\rho,\varepsilon}(vp).$$

Thus it remains to show that

$$\int \frac{v}{\rho} pR \frac{dx}{h} = \sum_{k=1}^{\nu^*} \frac{(\varepsilon_k - 1)}{(\nu_k - 1)!} \left(\frac{vpR}{\rho_k \sqrt{H}} \right)^{(\nu_k - 1)} (w_k).$$

Partial fraction expansion gives, for $v \in \mathbb{P}_{\nu-1}$,

$$\frac{v(x)}{\rho(x)} = \sum_{k=1}^{\nu^*} \sum_{j=1}^{\nu_k} \frac{A_{k,j}}{(x - w_k)^j}$$

where

$$A_{k,j} = \left(\frac{v}{\rho_k} \right)^{(\nu_k - j)} (w_k)/(\nu_k - j)!. \tag{4.5}$$

Since by Lemma 1(a) and (2.2)

$$\int \frac{(Rp)(x)}{x - z} \frac{dx}{h(x)} = \int \frac{(Rp)(x) - (Rp)(z)}{x - z} \frac{dx}{h(x)} + (Rp)(z) \int \frac{1}{x - z} \frac{dx}{h(x)}$$

$$= q(z) - \frac{(Rp)(z)}{\sqrt{H(z)}},$$

we deduce that for $z \in \mathbb{C} \setminus E_l$

$$j! \int \frac{(Rp)(x)}{(x - z)^{j+1}} \frac{dx}{h(x)} = q^{(j)}(z) - \left(\frac{Rp}{\sqrt{H}} \right)^{(j)} (z). \tag{4.6}$$

Taking into account the fact that p and q have no common zero, we obtain from (2.1) that

$$q^{(j)}(w_k) = \varepsilon_k \left(\frac{Rp}{\sqrt{H}} \right)^{(j)} (w_k) \quad \text{for } j = 0, \cdots, \nu_k - 1.$$

Thus, with the help of (4.5), (4.6), and Leibniz's formula, we get

$$\int \frac{v(x)}{\rho(x)} (Rp)(x) \frac{dx}{h(x)} = \sum_{k=1}^{\nu^*} \sum_{j=1}^{\nu_k} A_{k,j} \int \frac{(Rp)(x)}{(x - w_k)^j} \frac{dx}{h(x)}$$

$$= \sum_{k=1}^{\nu^*} \sum_{j=1}^{\nu_k} A_{k,j} (\varepsilon_k - 1) \left(\frac{Rp}{\sqrt{H}} \right)^{(j-1)} (w_k)/(j-1)!$$

$$= \sum_{k=1}^{\nu^*} (\varepsilon_k - 1) \left(\frac{vRp}{\rho_k \sqrt{H}} \right)^{(\nu_k - 1)} (w_k),$$

which is the desired result.

Part (b) can be demonstrated analogously.

Part (d) follows immediately from the relation

$$\int \frac{(Rp)(z)-(Rp)(x)}{z-x}\frac{dx}{h(x)} = p(z)\int \frac{R(z)-R(x)}{z-x}\frac{dx}{h(x)} + \int \frac{p(z)-p(x)}{z-x}\frac{R(x)}{h(x)}\,dx,$$

part (c), and Lemma 1(c).

Proof of Corollary 1. Let us consider the simplest case:

$$\rho = \tilde{\rho}(x - w_{k^*}).$$

Observing that

$$\tilde{\rho}_k(x) = \begin{cases} \rho_k(x)/(x-w_{k^*}) & \text{for } k \neq k^*, \\ \rho_{k^*}(x) & \text{for } k = k^*, \end{cases}$$

we get, since by Leibniz's formula for $\nu_{k^*} > 1$

$$\frac{(1-\varepsilon_{k^*})}{(\nu_{k^*}-1)!}\left(\frac{(x-w_{k^*})f}{\rho_{k^*}}\right)^{(\nu_{k^*}-1)}(w_{k^*}) = \frac{(1-\varepsilon_{k^*})}{(\nu_{k^*}-2)!}\left(\frac{f}{\rho_{k^*}}\right)^{(\nu_{k^*}-2)}(w_{k^*}),$$

that for $f \in \mathbb{P}$

$$\Psi_{R,\rho,\varepsilon}((x-w_{k^*})f) = \Psi_{R,\tilde{\rho},\tilde{\varepsilon}}(f).$$

Since, by Theorem 1,

$$\Psi_{R,\rho,\varepsilon}((x-w_{k^*})tp) = 0 \quad \text{for all } t \in \mathbb{P}_{n+l-3}$$

the assertion is proved for the simplest case. The corollary now follows by induction arguments.

Proof of Corollary 2. Since

$$\frac{q(z)}{p(z)} = \frac{1}{\sqrt{H(z)}} - \frac{1}{p(z)}\int \frac{p(x)}{z-x}\frac{dx}{h(x)} = \frac{1}{\sqrt{H(z)}} + O(z^{-(n+1)})$$

it follows that

$$p^2 - Hq^2 = O(z^{n+l-1}),$$

which, in view of Theorem 1, proves the assertion.

Proof of Theorem 3. (a) Since Y satisfies (2.4) and the relation

$$\frac{R(z)/\sqrt{H(z)} - Y(z)}{\rho(z)} = O(z^{-1}),$$

we deduce that

$$G(z) := \frac{R(z)/\sqrt{H(z)} - Y(z)}{\rho(z)} - \sum_{k=1}^{\nu^*}\sum_{j=1}^{\nu_k}\frac{A_{k,j}}{(z-w_k)^j}$$

where

$$(\nu_k - j)!A_{k,j} = [(R/\sqrt{H} - Y)/\rho_k]^{(\nu_k - j)}(w_k) \tag{4.7}$$

is analytic on $\mathbb{C}\backslash E_l$ and of order $O(z^{-1})$. Using Theorem 4.1 and p. 495 of [16], we get that

$$G(z)=\int \frac{1}{z-x}\frac{R(x)}{\rho(x)h(x)}\,dx. \tag{4.8}$$

Since, by Lemma 4 for sufficiently large $|z|$,

$$L_{R,\rho,\varepsilon}\left(\frac{1}{z-x}\right)=\sum_{k=1}^{\nu^*}(1-\varepsilon_k)\sum_{j=1}^{\nu_k}\frac{(R/\rho_k\sqrt{H})^{(\nu_k-j)}(w_k)}{(\nu_k-j)!(z-w_k)^j}$$

we obtain from (4.7) and (4.8) that ($|z|$ sufficiently large)

$$\begin{aligned}\frac{R(z)/\sqrt{H(z)}-Y(z)}{\rho(z)}&=\int\frac{1}{z-x}\frac{R(x)}{\rho(x)h(x)}\,dx+L_{R,\rho,\varepsilon}\left(\frac{1}{z-x}\right)\\&=\Psi_{R,\rho,\varepsilon}\left(\frac{1}{z-x}\right).\end{aligned}$$

Hence, by the orthogonality property of p,

$$\frac{p^{[1]}(z)}{p(z)}=\frac{R(z)/\sqrt{H(z)}-Y(z)}{\rho(z)}+O(z^{-(i_n+i_{n+1})}) \tag{4.9}$$

from which it follows that

$$\sqrt{H(z)}(Yp+\rho p^{[1]})(z)=(Rp)(z)+O(z^{l+\nu-i_{n+1}}). \tag{4.10}$$

Squaring this relation, we obtain that for $i_{n+1}+i_n+r\geqq\nu+l+1$

$$S(z)(Yp+\rho p^{[1]})^2(z)=Rp^2(z)+O(z^{l+\nu+i_n-i_{n+1}}).$$

Since Y satisfies (2.4), it follows that

$$Rp^2-S(Yp+\rho p^{[1]})^2=\rho g$$

where $g\in\mathbb{P}_{i_n+l-i_{n+1}}$.

Observing that the $O(\ )$ part of (4.9) is of the form

$$O(z^{-(i_n+i_{n+1})})=(-\Psi_{R,\rho,\varepsilon}(x^{i_{n+1}-1}p_{i_n}))z^{-(i_n+i_{n+1})}+O(z^{-(i_n+i_{n+1}+1)}),$$

we find that g has the given leading coefficient.

Part (c) follows immediately from Theorem 1 and Corollary 1.

(b) In view of Lemma 2 and the remark at the end of § 3, we can choose an $n\in\mathbb{N}$ such that $i_{n+1}\geqq\nu+1$. Then on the one hand, setting $\Psi_{R,\rho,\varepsilon}=\Psi$, we get

$$\begin{aligned}\Psi\left(\frac{\rho(z)p(z)-\rho(x)p(x)}{z-x}\right)&=\rho(z)\Psi\left(\frac{p(z)-p(x)}{z-x}\right)+\Psi\left(\frac{\rho(z)-\rho(x)}{z-x}p(x)\right)\\&=\rho(z)p^{[1]}(z)\end{aligned}$$

and on the other hand,

$$\begin{aligned}\Psi\left(\frac{\rho(z)p(z)-\rho(x)p(x)}{z-x}\right)&=p(z)\Psi\left(\frac{\rho(z)-\rho(x)}{z-x}\right)+\Psi\left(\frac{p(z)-p(x)}{z-x}\rho(x)\right)\\&=p(z)\Psi\left(\frac{\rho(z)-\rho(x)}{z-x}\right)+\int\frac{p(z)-p(x)}{z-x}\frac{R(x)}{h(x)}\,dx\\&=p(z)\left[\Psi\left(\frac{\rho(z)-\rho(x)}{z-x}\right)-\int\frac{R(z)-R(x)}{z-x}\frac{dx}{h(x)}\right]\\&\qquad+\int\frac{(Rp)(z)-(Rp)(x)}{z-x}\frac{dx}{h(x)},\end{aligned}$$

which, with the help of (a) and Theorem 1(c), gives the assertion.

Proof of Corollary 3. (a) Since Rp and q have no common zero and since

$$(Rp)^2 - Hq^2 = R\rho g,$$

it follows that $N = R\rho g > 0$ on int (E_l). Differentiating the relation

$$(RpN^{-1/2})^2 - H(qN^{-1/2})^2 = 1, \tag{4.11}$$

we get that at the zeros of q

$$(RpN^{-1/2})' = N^{-3/2}[-N'Rp/2 + (Rp)'N] = 0. \tag{4.12}$$

Thus there is a $\tilde{u} \in \mathbb{P}_{r+l+\nu+\mu-1}$ such that

$$\tilde{u}q = -N'Rp/2 + (Rp)'N,$$

which proves (a).

(b) From (4.11), (4.12), and (a) it follows that

$$y^2 + G(x)(y')^2 = 1.$$

Differentiating this relation, we have part (b).

5. Proofs of Theorems 4 and 5 and Corollaries 4 and 5. The following lemma plays a crucial role in what follows.

LEMMA 6. *Let*

$$R = AB, \quad S = CD \quad \text{and} \quad \tilde{R} = AD, \quad \tilde{S} = CB \tag{5.1}$$

where $A, B, C, D \in \mathbb{P}$, $RS = \tilde{R}\tilde{S} = H$ *and* B *and* D *have no common zero. Furthermore, assume that the polynomials* e_n, f_m, g_j, h_k *have the following properties*:

$$Re_n^2 - Sf_m^2 = \rho_\nu \tag{5.2}$$

where $2n + r > \nu$, e_n *and* f_m *have no common zero, and at the zeros* w_k *of* ρ_ν

$$Re_n(w_\kappa) = \varepsilon_k \sqrt{H(w_k)} f_m(w_\kappa) \quad \text{for } \kappa = 1, \cdots, \nu^*, \tag{5.3}$$

$$\tilde{R}g_j^2 - \tilde{S}h_k^2 = \tilde{\rho}_\mu \tag{5.4}$$

where $2j + \tilde{r} > \mu$, g_j *and* h_k *have no common zero, and at the zeros* $\tilde{w}_\kappa$ *of* $\tilde{\rho}_\mu$

$$\tilde{R}g_j(\tilde{w}_\kappa) = \tilde{\varepsilon}_\kappa \sqrt{H(\tilde{w}_\kappa)} h_k(\tilde{w}_\kappa) \quad \text{for } \kappa = 1, \cdots, \mu^*. \tag{5.5}$$

Then the following propositions hold:

(a) $BD(Ae_ng_j \pm Cf_mh_k)^2 - AC(Df_mg_j \pm Be_nh_k)^2 = \rho_\nu \tilde{\rho}_\mu$, *where at the zeros* w_κ *of* ρ_ν

$$BD(Ae_ng_j \pm Cf_mh_k)(w_\kappa) = \varepsilon_\kappa \sqrt{H(w_k)} (Df_mg_j \pm Be_nh_k)(w_\kappa),$$

and at the zeros $\tilde{w}_\kappa$ *of* $\tilde{\rho}_\mu$

$$BD(Ae_ng_j \pm Cf_mh_k)(\tilde{w}_\kappa) = \pm \tilde{\varepsilon}_\kappa \sqrt{H(\tilde{w}_\kappa)} (Df_mg_j \pm Be_nh_k)(\tilde{w}_\kappa).$$

(b) *If* $2n + \mu + (r - \tilde{r}) > 2j + \nu$, *then* $Ae_ng_j - Cf_mh_k$, *respectively,* $Df_mg_j - Be_nh_k$, *is of degree* $n - j + \mu - \partial D$, *respectively,* $m - j + \mu - \partial A$, *and both polynomials have leading coefficient* $\tilde{K}$, *where* $\tilde{K}$ *is the leading coefficient of* $\tilde{\rho}_\mu$.

Proof. (a) Simple calculation gives

$$BD(Ae_ng_j \pm Cf_mh_k)^2 - AC(Df_mg_j \pm Be_nh_k)^2 = (\tilde{R}g_j^2 - \tilde{S}h_k^2)(Re_n^2 - Sf_m^2) = \rho_\nu \tilde{\rho}_\mu.$$

Since, in view of (5.1) and (5.3) at the zeros w_κ of ρ_ν,

$$BDAe_ng_j = \varepsilon_\kappa \sqrt{H} f_m Dg_j \quad \text{and} \quad BDCf_mh_k = \varepsilon_\kappa \sqrt{H} e_n Bh_k,$$

and in view of (5.1) and (5.5),

$$BDAe_n g_j = \tilde{\varepsilon}_\kappa \sqrt{H}\, h_k Be_n \quad \text{and} \quad BDCf_m h_k = \tilde{\varepsilon}_\kappa \sqrt{H}\, g_j Df_m,$$

part (a) follows.

(b) From (5.2), respectively, (5.4), we derive that for $2n+r>\nu$

$$(f_m\sqrt{H})(z) = (Re_n)(z) - (K/2)z^{-(n-\nu)} + O(z^{-(n-\nu+1)}), \tag{5.6}$$

respectively, that for $2j+\tilde{r}>\mu$

$$(h_k\sqrt{H})(z) = (\tilde{R}g_j)(z) - (\tilde{K}/2)z^{-(j-\mu)} + O(z^{-(j-\mu+1)}), \tag{5.7}$$

where K, respectively, $\tilde{K}$, denotes the leading coefficient of ρ_ν, respectively, $\tilde{\rho}_\mu$.

Multiplying (5.6) and (5.7) and observing that

$$H = ABCD, \quad R\tilde{R} = A^2BD, \quad n+j+\partial A = m+k+\partial C, \quad m+j+\partial D = n+k+\partial B,$$

we get

$$Ae_n g_j - Cf_m h_k = (\tilde{K}/2)z^{n+\mu+\partial A-(j+\tilde{r})} + \text{lower terms in } z,$$

respectively,

$$Df_m g_j - Be_n h_k = (K/2)z^{m+\mu+\partial D-(j+\tilde{r})} + \text{lower terms in } z,$$

which gives part (b).

Proof of Theorem 4. (a) By Theorem 3 we have for $i_{n+1}+i_n+r>\nu+l$

$$Rp_{i_n}^2 - Sq_{i_m}^2 = \rho_\nu g_{(i_n)}$$

where $\partial g_{(i_n)} = l+i_n-i_{n+1}$, and at the zeros w_κ of ρ_ν

$$Rp_{i_n}(w_\kappa) = \varepsilon_\kappa \sqrt{H(w_\kappa)}\, q_{i_m}(w_\kappa) \quad \text{for } \kappa = 1, \cdots, \nu^*.$$

From Lemma 6 we obtain that

$$RP_{i_n+j}^2 - SQ_{i_m+j}^2 = \rho_\nu \tilde{\rho}_\mu g_{(i_n)}$$

and that

$$RP_{i_n+j}(w_\kappa) = \varepsilon_\kappa \sqrt{H(w_k)}\, Q_{i_m+j}(w_\kappa) \quad \text{and} \quad RP_{i_n+j}(\tilde{w}_\kappa) = \tilde{\varepsilon}_\kappa \sqrt{H(\tilde{w}_\kappa)}\, Q_{i_m+j}(\tilde{w}_\kappa)$$

for $\kappa = 1, \cdots, \nu^*$, respectively, $\kappa = 1, \cdots, \mu^*$.

Applying Theorem 1 and Corollary 1 gives part (a). Part (b) can be proved analogously. Part (c) follows immediately from the definition of $P_{i_n+j}, \cdots$, and Lemma 2(c).

Proof of Corollary 4. Let (l_i) be the sequence of basic integers with respect to $\Psi_{R,\rho_\nu\bar{\rho}_{\mu-\tau},\delta}$ and let $i^* \in \mathbb{N}_0$ be such that $l_{i^*} \leqq n+j \leqq l_{i^*+1}-1$. Then $gp_{l_{i^*}}$ can be represented in the form

$$gp_{l_{i^*}} = \sum_{\kappa=0}^{\tau+l_{i^*}} c_{\kappa,n}P_\kappa = \sum_{\kappa=l_{i^*+1}-1}^{\tau+l_{i^*}} c_{\kappa,n}P_\kappa$$

where the last equality follows from Theorem 4(a) and the fact that $gp_{l_{i^*}} \perp \mathbb{P}_{l_{i^*+1}-2}$ with respect to $\Psi_{R,\rho_\nu\bar{\rho}_{\mu-\tau}g,\varepsilon}$. Hence the corollary is proved for $\bar{P}_{n+j}$.

Concerning $\bar{Q}_{m+j}$, we observe that in view of (2.6) and (2.7)

$$Rv_k \sum_{\kappa=0}^{\tau} c_{\kappa,n}P_{n+j+\kappa} = u_j \sum_{\kappa=0}^{\tau} c_{\kappa,n}Q_{m+j+\kappa} - \bar{\rho}_{\mu-\tau}g \sum_{\kappa=0}^{\tau} c_{\kappa,n}q_{m+\kappa}.$$

Since, by (2.6), u_j does not vanish at the zeros of g, we obtain that $\bar{Q}_{m+j} \in \mathbb{P}_{m+j}$. The assertion now follows from Theorem 4(a).

Proof of Corollary 5. First let us demonstrate that the assertion holds for $\mu = 0$. Since t_ν has all zeros in the open unit disk it follows that

$$(\operatorname{Re}\{t_\nu(e^{i\varphi})\})^2 + \sin^2\varphi(\operatorname{Im}\{t_\nu(e^{i\varphi})\}/\sin\varphi)^2 = \rho_\nu(\cos\varphi) > 0, \tag{5.8}$$

from which we get by the orthogonality property (Theorem 1) that the polynomials $\operatorname{Re}\{t_\nu(e^{i\varphi})\}$ and $\operatorname{Im}\{t_\nu(e^{i\varphi})\}/\sin\varphi$, $x = \cos\varphi$, have ν, respectively, $\nu - 1$, simple zeros in $(-1, +1)$ that, because of (5.8), separate each other. Using the well-known fact that $T_n(x) = \operatorname{Re} e^{in\varphi}$, respectively, $U_{n-1}(x) = \operatorname{Im} e^{in\varphi}/\sin\varphi$, $n \in \mathbb{N}$, is orthogonal with respect to $1/\sqrt{1-x^2}$, respectively, $\sqrt{1-x^2}$, on $[-1, +1]$ we obtain from Theorem 4(a) combined with Theorem 2 that

$$\operatorname{Re}\{t_\nu(e^{i\varphi})\} \operatorname{Re} e^{in\varphi} + (x^2-1)\frac{\operatorname{Im}\{t_\nu(e^{i\varphi})\}}{\sin\varphi}\frac{\operatorname{Im} e^{in\varphi}}{\sin\varphi} = \operatorname{Re}\{e^{in\varphi}t_\nu(e^{i\varphi})\},$$

respectively,

$$\operatorname{Re}\{t_\nu(e^{i\varphi})\}\frac{\operatorname{Im} e^{in\varphi}}{\sin\varphi} + \frac{\operatorname{Im}\{t_\nu(e^{i\varphi})\}}{\sin\varphi}\operatorname{Re} e^{in\varphi} = \frac{\operatorname{Im}\{e^{in\varphi}t_\nu(e^{i\varphi})\}}{\sin\varphi},$$

is orthogonal with respect to $\Psi_{\rho_\nu(1,1,\cdots,1)}$, respectively, $\Psi_{x^2-1,\rho_\nu(1,1,\cdots,1)}$. Hence the assertion is proved for $\mu = 0$.

Using the relations

$$(\operatorname{Re}\{\tilde{t}_\mu(e^{-i\varphi})\})^2 + \sin^2\varphi\left(\frac{\operatorname{Im}\tilde{t}_\mu(e^{-i\varphi})}{\sin\varphi}\right)^2 = \tilde{\rho}_\mu(\cos\varphi)$$

and, setting $\kappa = \mu + \nu$,

$$\operatorname{Re}\{\tilde{t}_\mu(e^{-i\varphi})\} \operatorname{Re}\{e^{i(n-\kappa)\varphi}t_\nu(e^{i\varphi})\} + \sin^2\varphi\frac{\operatorname{Im}\{\tilde{t}_\mu(e^{-i\varphi})\}}{\sin\varphi}\frac{\operatorname{Im}\{e^{i(n-\kappa)\varphi}t_\nu(e^{i\varphi})\}}{\sin\varphi}$$
$$= \operatorname{Re}\{e^{i(n-\kappa)\varphi}t_\nu(e^{i\varphi})\tilde{t}_\mu(e^{i\varphi})\},$$

we get the assertion from the first part of the proof and Theorem 4(b).

Proof of Theorem 5. (a) Using (2.5) and relation (10) of [9], we get that

$$q_{n+k}p_{j-1} - p_{n+j}q_{k-1} = \rho_\nu(p^{[1]}_{n+j}p_{j-1} - p^{[1]}_{j-1}p_{n+j})$$
$$= \rho_\nu\left(\prod_{i=1}^{j}\lambda_i\right)p_n^{(j)},$$

which proves (a).

(b), (c). In view of Lemma 6 and (2.9) we have for $j \in \mathbb{N}_0$

$$(Rp_{n+j}p_{j-1} - Sq_{n+k}q_{k-1})^2 - H(q_{n+k}p_{j-1} - p_{n+j}q_{k-1})^2 = \rho_\nu^2 g_{(j-1)}g_{(n+j)}, \tag{5.9}$$

with

$$(Rp_{n+j}p_{j-1} - Sq_{n+k}q_{k-1})(z_{\kappa,j-1})$$
$$= -\delta_{\kappa,j-1}(\sqrt{H}(q_{n+k}p_{j-1} - p_{n+j}q_{k-1}))(z_{\kappa,j-1}) \quad \text{for } \kappa = 1, \cdots, l^* - 1. \tag{5.10}$$

From (5.9) and (a) we obtain that $\partial C^{(j)}_{n+l} = n + l$. Since, in view of (2.5), $(p_{n+j})_{n\in\mathbb{N}_0}$ and $(q_{n+k})_{n\in\mathbb{N}_0}$ satisfy the recurrence relation

$$y_n = (x - \alpha_{n+j})y_{n-1} - \lambda_{n+j}y_{n-2},$$

part (b) follows. Part (c) follows immediately from (5.9), (5.10), and Theorem 1.

6. Proofs of Theorems 2 and 6. In the following we put

$$p = p_n, \quad q = q_m, \quad \rho = \rho_\nu.$$

Proof of Theorem 2. Set

$$u(x) = \prod_{j=2}^{l} (x - a_{2j-1}).$$

Necessity. From Theorem 1 and Lemma 5 it follows that all zeros of p and q are simple and lie in int (E_l). Now let

$$x_1 < x_2 < \cdots < x_n$$

denote the zeros of p. Since, by (2.1'), $\rho > 0$ on int (E_l), we get that $u/\rho h > 0$ on int (E_l). Taking into account that, in view of Theorem 1, $p \perp \mathbb{P}_{n-1}$ with respect to $u/\rho h$, we obtain by Gaussian quadrature formula that

$$\int \frac{tu}{h\rho}\, dx = \sum_{j=1}^{n} \lambda_j t(x_j) \quad \text{for all } t \in \mathbb{P}_{2n-1'}$$

where $\lambda_j \in \mathbb{R}^+$. Hence

$$\int tu \frac{dx}{h} = \sum_{j=1}^{n} \lambda_j \rho(x_j) t(x_j) \quad \text{for all } t \in \mathbb{P}_{2n-1-\nu}. \tag{6.1}$$

Since, by Theorem 1(c) and Corollary 1,

$$u(z)q(z) = \int \frac{u(z)p(z) - u(x)p(x)}{z - x} \frac{dx}{h(x)},$$

we get with the help of (6.1) that at a zero x_{j^*} of p

$$u(x_{j^*})q(x_{j^*}) = \lambda_{j^*}\rho(x_{j^*})p'(x_{j^*}),$$

which implies that uq and p have strictly interlacing zeros. Thus the necessity part is proved.

Sufficiency. Since all zeros of p and q are simple and are in int (E_l), and since the zeros of p and uq interlace strictly, we get that p and $(x - a_1)uq$ have the same number of sign changes in each interval $[a_{2j-1}, a_{2j})$, $j = 1, \cdots, l$, which implies that

$$\operatorname{sgn} p(x) = \operatorname{sgn} \prod_{j=1}^{l} (x - a_{2j-1}) \operatorname{sgn} q(x) \quad \text{for } x \in \mathbb{R} \setminus E_l.$$

Hence we obtain from (2.1) and Lemma 1(b) that at the real zeros w_k of ρ

$$p(w_k) = \sqrt{H(w_k)}\, q(w_k).$$

Thus it remains to show that at the complex zeros w_j of ρ

$$p(w_j) = \sqrt{H(w_j)}\, q(w_j). \tag{6.2}$$

Since on the one hand, by partial fraction expansion,

$$\frac{(uq)(z)}{p(z)} = \sum_{j=1}^{n} \frac{\lambda_j}{z - x_j}, \qquad \lambda_j \in \mathbb{R}^+$$

and on the other hand, by Lemma 1(a),

$$\frac{u(z)}{\sqrt{H(z)}} = \int \frac{u(x)}{z - x} \frac{dx}{h(x)},$$

we obtain with the help of the relations

$$\operatorname{Im}\{1/(z-x_j)\} \gtrless 0 \quad \text{for } \operatorname{Im} z \gtrless 0, \qquad u/h > 0 \quad \text{on int}(E_l)$$

and that

$$\operatorname{sgn} \operatorname{Im}\{(uq)(z)/p(z)\} = \operatorname{sgn} \operatorname{Im}\{u(z)/\sqrt{H(z)}\} \quad \text{for } z \in \mathbb{C}\backslash\mathbb{R},$$

which implies that (6.2) holds at the zeros of ρ.

Proof of Theorem 6. Since $1/h\rho > 0$ on E_l it follows that ρ_1 has in each interval (a_{2j}, a_{2j+1}), $j = 1, \cdots, l-1$, exactly one zero more than ρ_2. Hence

$$\partial\rho_1 = \partial\rho_2 + l - 1 \quad \text{and} \quad \partial(\rho_1 q) = \partial(\rho_2 p) - 1.$$

Furthermore, we deduce that the zeros of ρ_1 and $\prod_{j=2}^{l-1}(x-a_{2j})\rho_2$ interlace strictly, which gives, in conjunction with the relation

$$\operatorname{sgn}\sqrt{H(x)} = \begin{cases} -\operatorname{sgn}\prod_{j=2}^{l-1}(x-a_{2j}) & \text{for } x \in [a_1, a_{2l}]\backslash E_l, \\ \operatorname{sgn}\prod_{j=2}^{l-1}(x-a_{2j}) & \text{for } x \in \mathbb{R}\backslash[a_1, a_{2l}], \end{cases} \tag{6.3}$$

that at the zeros $w_{k,1}$ of ρ_1

$$\operatorname{sgn}\sqrt{H(w_{k,1})}\operatorname{sgn}\rho_2(w_{k,1}) = -\operatorname{sgn}\rho_1'(w_{k,1}) \tag{6.4}$$

and at the zeros $w_{k,2}$ of ρ_2

$$\operatorname{sgn}\rho_1(w_{k,2}) = \operatorname{sgn}\sqrt{H(w_{k,2})}\operatorname{sgn}\rho_2'(w_{k,2}). \tag{6.5}$$

Necessity. Since $1/\rho h > 0$ it is well known (see, e.g., [7]), that the zeros of p and $p^{[1]}$ interlace strictly. Hence we get from Theorem 3(b) that at the zeros x_k of p

$$\operatorname{sgn} q(x_k) = \operatorname{sgn}\rho(x_k)\operatorname{sgn}p^{[1]}(x_k) = \operatorname{sgn}\rho(x_k)\operatorname{sgn}p'(x_k),$$

which implies immediately that

$$\operatorname{sgn}(\rho_1 q)(x_k) = \operatorname{sgn}(\rho_2 p)'(x_k).$$

Since, by Theorem 1,

$$p(w_k) = \sqrt{H(w_k)}q(w_k) \quad \text{for } k = 1, \cdots, \mu,$$

we obtain from (6.5) that at the zeros $w_{k,2}$ of ρ_2

$$\operatorname{sgn}(\rho_1 q)(w_{k,2}) = \operatorname{sgn}(\rho_2 p)'(w_{k,2}),$$

which proves the necessity part.

Sufficiency. From the interlacing property of $\rho_1 q$ and $\rho_2 p$ it follows that at the zeros $w_{k,2}$ of ρ_2

$$\operatorname{sgn}(\rho_1 q)(w_{k,2}) = \operatorname{sgn}\rho_2'(w_{k,2})\operatorname{sgn}p(w_{k,2})$$

and at the zeros $w_{k,1}$ of ρ_1

$$\operatorname{sgn}(\rho_2 p)(w_{k,1}) = -\operatorname{sgn}\rho_1'(w_{k,1})\operatorname{sgn}q(w_{k,1}),$$

which implies, by (6.4) and (6.5), that

$$\operatorname{sgn}p(w_k) = \operatorname{sgn}\sqrt{H(w_k)}\operatorname{sgn}q(w_k) \quad \text{for } k = 1, \cdots, \mu. \tag{6.6}$$

Using the fact that

$$\operatorname{sgn} q(x) = \operatorname{sgn} \sqrt{H(x)} \operatorname{sgn} p(x) \quad \text{for } x \in \mathbb{R}\backslash[a_1, a_{2l}],$$

we find with the help of (2.10) that at all real zeros w of ρ

$$p(w) = \sqrt{H(w)}\, q(w). \tag{6.7}$$

Since p and q have at most one zero in each interval (a_{2j}, a_{2j+1}), $j = 1, \cdots, l-1$, and since the zeros of $\rho_1 q$ and $\rho_2 p$ are strictly interlacing, we can choose exactly one zero $w_{k_j,1}$ of ρ_1 from each interval (a_{2j}, a_{2j+1}), $j = 1, \cdots, l-1$, such that the zeros of $\tilde{\rho}_1 q$ and p interlace strictly, where

$$\tilde{\rho}_1(x) = \prod_{j=1}^{l-1} (x - w_{k_j,1}).$$

If we proceed as in the proof of the sufficiency part of Theorem 2, it follows that (6.7) holds also at the complex conjugate zeros of ρ, which proves the theorem.

Acknowledgments. The author thanks the referees for valuable comments, for pointing out an error in the original version of the manuscript, and for bringing papers [26], [28], and [29] to his attention.

REFERENCES

[1] N. I. ACHIESER, *Vorlesungen über Approximationstheorie*, Akademie-Verlag, Berlin, 1953.

[2] ———, *Orthogonal polynomials on several intervals*, Soviet. Math. Dokl., 1 (1960), pp. 989-992.

[3] N. I. ACHIESER AND YU. YA. TOMCUK, *On the theory of orthogonal polynomials over several intervals*, Soviet. Math. Dokl., 2 (1961), pp. 687-690.

[4] F. V. ATKINSON AND W. N. EVERITT, *Orthogonal polynomials which satisfy second order differential equations*, in E. B. Christoffel, The Influence of His Work on Mathematics and the Physical Sciences, P. L. Butzer and F. Fehér, eds., Birkhäuser, Boston, 1981, pp. 173-181.

[5] S. N. BERNSTEIN, *Sur une classe de polynomes orthogonaux*, Comm. Kharkov Math. Soc., 4 (1930), pp. 79-93.

[6] P. L. CHEBYSHEV, *Sur les quéstions de minima qui se rattachent à la représentation approximative des fonctions*, in Collected Works, Vol. I, Chelsea, New York, 19xx, pp. 273-378.

[7] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[8] ———, *On co-recursive orthogonal polynomials*, Proc. Amer. Math. Soc., 8 (1957), pp. 899-905.

[9] D. J. DICKINSON, *On certain polynomials associated with orthogonal polynomials*, Boll. Un. Mat. Ital. (3), 13 (1958), pp. 116-124.

[10] J. L. GAMMEL AND J. NUTTALL, *Note on generalized Jacobi polynomials*, in The Riemann Problem, Complete Integrability and Arithmetic Applications, Lecture Notes in Math. 925, Springer-Verlag, Berlin, New York, 1982, pp. 258-270.

[11] J. S. GERONIMO AND W. VAN ASSCHE, *Orthogonal polynomials with asymptotically periodic recurrence coefficients*, J. Approx. Theory, 46 (1986), pp. 251-283.

[12] ———, *Orthogonal polynomials on several intervals via a polynomial mapping*, Trans. Amer. Math. Soc., 308 (1988), pp. 559-581.

[13] YA. L. GERONIMUS, *On some finite difference equations and corresponding systems of orthogonal polynomials*, Mem. Math. Sect. Fac. Phys. Kharkov State Univ. and Kharkov Math. Soc., 25 (1975), pp. 81-100.

[14] W. HAHN, *Uber Differentialgleichungen für Orthogonalpolynome*, Monatsh. Math., 95 (1983), pp. 269-274.

[15] M. E. H. ISMAIL, *On sieved orthogonal polynomials* III, *Orthogonality on several intervals*, Trans. Amer. Math. Soc., 294 (1986), pp. 89-111.

[16] N. LEVINSON, *Simplified treatment of integrals of Cauchy type, the Hilbert problem and singular integral equations*, SIAM Rev., 7 (1965), pp. 474-502.

[17] A. MAGNUS, *Recurrence coefficients for orthogonal polynomials on connected and nonconnected sets*, in Padé Approximation and its Applications, Lecture Notes in Math. 765, Springer-Verlag, Berlin, New York, 1979, pp. 150-171.

[18] J. NUTTALL AND S. R. SINGH, *Orthogonal polynomials and Padé approximants associated with a system of arcs*, J. Approx. Theory, 21 (1977), pp. 1-42.

[19] J. NUTTALL, *Asymptotics of diagonal Hermite-Padé polynomials*, J. Approx. Theory, 42 (1984), pp. 299-386.

[20] F. PEHERSTORFER, *Orthogonal polynomials in L^1-approximation*, J. Approx. Theory, 52 (1988), pp. 241-268.

[21] ———, *On Tchebycheff polynomials on disjoint intervals*, in Haar Memorial Conference, J. Szabados and K. Tandori, eds., North-Holland, Amsterdam, New York, 1988, pp. 737-751.

[22] ———, *Orthogonal- and Chebyshev polynomials on two intervals*, Acta. Math. Hungar. (1990), to appear.

[23] ———, *On Gauss quadrature formulas with equal weights*, Numer. Math., 52 (1988), pp. 317-327.

[24] J. RICE, *The Approximation of Functions.* II, Addison-Wesley, Reading, MA, 1969.

[25] G. SZEGÖ, *Orthogonal polynomials*, Amer. Math. Soc. Colloq. Publ. 23, Fourth edition, American Mathematical Society, Providence, RI, 1975.

[26] P. NEVAI, *A new class of orthogonal polynomials*, Proc. Amer. Math. Soc., 91 (1984), pp. 409-415.

[27] A. DRAUX, *Polynomes Orthogonaux Formels—Applications*, Lecture Notes in Math. 974, Springer-Verlag, Berlin, New York, 1983.

[28] J. DOMBROWSKI AND P. NEVAI, *Orthogonal polynomials, measures and recurrence relations*, SIAM J. Math. Anal., 17 (1986), pp. 752-759.

[29] J. M. COHEN AND A. R. TRENHOLME, *Orthogonal polynomials with a constant recursion formula and an application to harmonic analysis*, J. Funct. Anal., 59 (1984), pp. 175-184.

[30] C. BRÉZINSKI, *Padé-Type Approximation and General Orthogonal Polynomials*, Internat. Ser. Numer. Math., 50, Birkhäuser, Basel, 1980.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 483-509, March 1990
© 1990 Society for Industrial and Applied Mathematics
011

ZEROS OF TRANSFORMED POLYNOMIALS*

A. ISERLES† AND S. P. NØRSETT†

Abstract. Transformations that map polynomials with zeros in a certain interval into polynomials with zeros in another interval are considered here. By using the theory of bi-orthogonal polynomials, a general technique for the construction of such transformations is developed. Finally, a list of 16 different transformations formed by using the authors' technique is presented.

Transformations of this type have already been applied in numerical analysis, approximation theory, and real analysis.

Key words. orthogonal polynomials, hypergeometric functions, bi-orthogonal polynomials, q-hypergeometric functions, zeros of polynomials

AMS(MOS) subject classifications. primary 26C10; secondary 26C05, 33A35, 33A65, 42C05

1. Introduction.

Let D and E be two real intervals that need not be distinct. The theme of this paper is transformations that map polynomials whose zeros all lie in D into polynomials with all zeros in E.

Many nontrivial instances of such transformations are known ([Marden, 1966], [Pólya and Szegö, 1976]). Probably the most important are due to Pólya and Schur [1914]: Let the transformation T be defined as

$$T\left\{\sum_{k=0}^{m} p_k x^k\right\} = \sum_{k=0}^{m} \alpha_k p_k x^k, \tag{1.1}$$

where $\{\alpha_k\}_{k=0}^{\infty}$ is a given real sequence (called a *multiplier sequence*). Then, given that all the zeros of $p(x)=\Sigma p_k x^k$ are real, all the zeros of Tp will also be always real if and only if the function $f(z)=\sum_{k=0}^{\infty}(1/k!)\alpha_k z^k$ is of the form

$$f(z) = c\, e^{dz} z^n \prod_{l=1}^{N}\left(1+\frac{z}{z_l}\right), \tag{1.2}$$

where n is a nonnegative integer, N is either a nonnegative integer or infinity, $c\in\mathscr{R}$, $d\geqq 0$, $z_l>0$ for all $1\leqq l\leqq N$ and $\sum_{l=1}^{N} z_l^{-1}<\infty$. See also [Hille, 1962], [Pólya and Szegö, 1976], as well as the extension of these transformations for complex $D=E$ by Craven and Csordas [1983].

In the present paper we introduce a mechanism for generation of numerous transformations of similar type. It is based on the theory of bi-orthogonal polynomials that has been recently developed by the authors [Iserles and Nørsett, 1988]. This mechanism, as well as relevant elements of the theory of bi-orthogonal polynomials, is described in § 2. By using our mechanism and further algebraic manipulation, we introduce sixteen transformations in §§ 3-6. These transformations are listed below, with names ("Jacobi," "Laguerre," etc.) given for identification purposes. The reason for these names will become clear in the sequel. We use throughout the standard notation for Pochhammer symbols $(a)_m$:

$$(a)_0 := 1, \qquad a\in\mathscr{C};$$
$$(a)_m := (a)_{m-1}(a+m-1), \qquad m\geqq 1,$$

* Received by the editors March 7, 1988; accepted for publication (in revised form) February 13, 1989.

† Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom. This paper was written while the author was visiting the Department of Mathematics, University of Arizona, Tucson, Arizona.

‡ Division of Mathematical Sciences, Norwegian Institute of Technology, Trondheim, Norway.

and for Gauss-Heine symbols $(a;\rho)_m$:

$$(a;\rho)_0 := 1, \qquad a, \rho \in \mathscr{C},$$
$$(a;\rho)_m := (a;\rho)_{m-1}(1-a\rho^{m-1}), \qquad m \geqq 1.$$

Zero-mapping transformations.

1. *Jacobi transformation.*

$$T\left\{\sum_{k=0}^{m} r_k(x)_k\right\} = \sum_{k=0}^{m} (-1)^k(-\alpha-m)_k x^k (1-x)^{m-k} r_k,$$
$$\alpha > 1, \quad D = (0,\infty), \quad E = (0,1).$$

2. *Laguerre transformation.*

$$T\left\{\sum_{k=0}^{m} r_k(x)_k\right\} = \sum_{k=0}^{m} r_k x^k, \qquad D = E = (0,\infty).$$

3. *Meixner transformation.*

$$\lambda > 0, \quad T\left\{\sum_{k=0}^{m} q_k x^k\right\} = \sum_{k=0}^{m} (-1)^k(-x)_k(\lambda+x)_{m-k}\lambda^k q_k, \quad D = (0,\lambda), \quad E = (0,\infty).$$

4. *Charlier transformation.*

$$T\left\{\sum_{k=0}^{m} q_k x^k\right\} = \sum_{k=0}^{m} (-1)^k(-x)_k q_k, \qquad D = E = (0,\infty).$$

5. *Krawtchouk transformation.*

$$\eta \geqq m, \quad T\left\{\sum_{k=0}^{m} q_k x^k\right\} = \sum_{k=0}^{m} (-x)_k(x-\eta)_{m-k} q_k, \quad D = E = (0,\infty).$$

6. *Wall transformation.*

$$T\left\{\sum_{k=0}^{m} s_k(x;\rho)_k\right\} = \sum_{k=0}^{m} (\rho^{m-k}\alpha;\rho)_k(\alpha x;\rho)_{m-k} s_k x^k,$$
$$\rho \in (0,1), \quad \alpha \in (-1,1), \quad D = E = (0,1).$$

7. *Llaw tranformation.*

$$\rho > 1, \quad T\left\{\sum_{k=0}^{m} (-x;\rho)_k s_k\right\} = \sum_{k=0}^{m} s_k x^k, \quad D = (0,\infty), \quad E = (1,\infty).$$

8. *q-Krawtchouk transformation.*

$$T\left\{\sum_{k=0}^{m} (-x\rho^N;\rho)_k s_k\right\} = \sum_{k=0}^{m} s_k(\rho^{m-N-k};\rho)_k(\rho^{-N}x;\rho)_{m-k} x^k,$$
$$N \geqq m, \quad \rho \in (0,1), \quad D = (0,\infty), \quad E = (0,1).$$

9. *Wigert transformation.*

$$T\left\{\sum_{k=0}^{m} t_k x^k(-\alpha\rho^{-m+1}x,\rho)_{m-k}\rho^{-(3/2)k^2}\right\} = \sum_{k=0}^{m} t_k x^k,$$
$$\alpha < 1, \quad \alpha \neq 0, \quad \rho \in (0,1), \quad D = E = (0,\infty).$$

10. Γ_1 *transformation.*

$$\lambda > 0, \quad T\left\{\sum_{k=0}^{m} r_k(x)_k\right\} = \sum_{k=0}^{m} (-1)^k(-x)_k\lambda^{m-k} r_k, \quad D = E = (0,\infty).$$

11. Γ_2 *transformation.*

$$a, b>0, \quad T\left\{\sum_{k=0}^{m} u_k(x-m)_k\right\} = \sum_{k=0}^{m} \frac{(-1)^k}{(a)_k(b)_k} u_{m-k}(-x)_k, \quad D=E=(0,\infty).$$

12. Γ_3 *transformation.*

$$T\left\{\sum_{k=0}^{m} r_k(x)_k\right\} = \sum_{k=0}^{m} (-1)^k(-m-\alpha+1)_k(-x)_k(x-N)_{m-k}r_k,$$

$$\alpha>0, \quad N\in\mathscr{Z}, \quad N\geqq m, \quad D=E=(0,\infty).$$

13. Γ_4 *transformation.*

$$T\left\{\sum_{k=0}^{m} r_k(x)_k\right\} = \sum_{k=0}^{m} (-1)^k(\beta-m)_k(-x)_k(x+\alpha)_{m-k}r_k,$$

$$\alpha, \beta>0, \quad \beta\notin\mathscr{Z}, \quad D=(0,\beta), \quad E=(0,\infty).$$

14. $q-\Gamma_1$ *transformation.*

$$\alpha, \rho\in(0,1), \quad T\left\{\sum_{k=0}^{m} w_k(\alpha x\rho^k;\rho)_{m-k}\right\} = \sum_{k=0}^{m} \frac{w_k x^k}{(\alpha;\rho)_k}, \quad D=E=(0,1).$$

15. $q-\Gamma_2$ *transformation.*

$$T\left\{\sum_{k=0}^{m} (-1)^k\rho^{kN-(k-1)k}(\rho^{-N};\rho)_k(\alpha;\rho)_k\left(-\frac{x}{\alpha}\rho^{N-m};\rho\right)_{m-k} v_k\right\} = \sum_{k=0}^{m} v_k \prod_{j=0}^{k-1}(x-\rho^j),$$

$$\alpha>\rho^{-N+1}, \quad \rho\in(0,1), \quad m\leqq N, \quad N\in\mathscr{Z}, \quad D=E=(0,1).$$

16. $q-\Gamma_3$ *transformation.*

$$T\left\{\sum_{k=0}^{m} (x;\rho)_k(x\rho^{\alpha+k+1};\rho)_{m-k}v_k\right\} = \sum_{k=0}^{m} \rho^{-(N+1+\alpha)k}(\rho^{-N+k};\rho)_{m-k}v_k \prod_{j=0}^{k-1}(x-\rho^j),$$

$$\alpha>0, \quad \rho\in(0,1), \quad m\leqq N, \quad N\in\mathscr{Z}, \quad D=E=(0,1).$$

An alternative technique for generating transformations with predictable behaviour of zeros can be developed from the work of Al-Salam and Ismail [1976] on *convolution-orthogonal polynomials*: Let ψ be a Laplace transform of a nonnegative function and assume that it is analytic (this assumption can be somewhat relaxed) and with nonzero derivatives at the origin. Then the transformation

$$T\left\{\sum_{k=0}^{m} q_k x^k\right\} = \sum_{k=0}^{m} \frac{q_k}{\psi^{(k)}(0)}(-x)_k \tag{1.3}$$

maps polynomials with real zeros into polynomials with real zeros. Full details of the proof and examples are reported in [Iserles, Nørsett, and Saff, 1988]. Note that a straightforward application of (1.3) generalises our Charlier transformation.

Some applications of results from the present paper feature in [Iserles and Saff, 1987]. Further mechanisms for generation of "zero-mapping" transformations are described in [Iserles and Saff, 1989] and [Iserles, Nørsett, and Saff, 1988].

2. A mechanism for generation of transformations. We commence this section by reviewing pertinent elements of the theory of bi-orthogonal polynomials from [Iserles and Nørsett, 1988].

Let $\pi_m[x]$ denote the set of all mth degree polynomials. Given a one-parametric family of distributions $\varphi(x, \mu)$, $x \in E$, $\mu \in D$, where D and E are intervals, we define the mth *bi-orthogonal polynomial* $p_m(x; \mu_1, \mu_2, \cdots, \mu_m)$ as a monic member of $\pi_m[x]$ that satisfies the conditions

$$\int_E p(x; \mu_1, \cdots, \mu_m)\, d\varphi(x, \mu_l) = 0, \qquad l = 1, 2, \cdots, m.$$

Here $\mu_1, \cdots, \mu_m$ are distinct points in D.

We denote the moments of φ by I_k, $k \geqq 0$:

$$I_k(\mu) := \int_E x^k\, d\varphi(x, \mu), \quad k \geqq 0, \quad \mu \in D,$$

and define the function $D_m(\mu_1, \cdots, \mu_m)$ by $D_0 \equiv 1$:

$$D_m(\mu_1, \cdots, \mu_m) := \det \begin{pmatrix} I_0(\mu_1), & I_1(\mu_1), & \cdots, & I_{m-1}(\mu_1) \\ \vdots & \vdots & & \vdots \\ I_0(\mu_m), & I_1(\mu_m), & \cdots, & I_{m-1}(\mu_m) \end{pmatrix}, \tag{2.1}$$

$$m \geqq 0, \quad \mu_1, \cdots, \mu_m \in D.$$

The mth *generator* of φ is defined as

$$H_m(\mu; \mu_1, \cdots, \mu_m) := \int_E p(x; \mu_1, \cdots, \mu_m)\, d\varphi(x, \mu), \qquad \mu \in D. \tag{2.2}$$

THEOREM 1. (a) *Bi-orthogonal polynomials exist and are unique for all* $m \geqq 1$ *and all distinct* $\mu_1, \cdots, \mu_m \in D$ *if and only if* $D_m(\mu_1, \cdots, \mu_m) \neq 0$ *for all* $m \geqq 1$ *and distinct* $\mu_1, \cdots, \mu_m \in D$.

(b) *If the last condition is satisfied then*

$$p_m(x; \mu_1, \cdots, \mu_m) = \frac{1}{D_m(\mu_1, \cdots, \mu_m)} \det \begin{pmatrix} I_0(\mu_1), & I_1(\mu_1), & \cdots, & I_m(\mu_1) \\ \vdots & \vdots & & \vdots \\ I_0(\mu_m), & I_1(\mu_m), & \cdots, & I_m(\mu_m) \\ 1, & x, & \cdots, & x^m \end{pmatrix}, \tag{2.3}$$

$$H_m(\mu; \mu_1, \cdots, \mu_m) = \frac{D_{m+1}(\mu_1, \cdots, \mu_m, \mu)}{D_m(\mu_1, \cdots, \mu_m)}.$$

Let φ be of the form $d\varphi(x, \mu) = \omega(x, \mu)\, d\alpha(x)$, where α is a distribution in x, independent of μ. We say that φ has the *interpolation property* if

$$E_m \begin{pmatrix} x_1, \cdots, x_m \\ \mu_1, \cdots, \mu_m \end{pmatrix} := \det \begin{pmatrix} \omega(x_1, \mu_1), & \cdots, & \omega(x_1, \mu_m) \\ \vdots & & \vdots \\ \omega(x_m, \mu_1), & \cdots, & \omega(x_m, \mu_m) \end{pmatrix} \neq 0$$

for all $m \geqq 1$, distinct $x_1, \cdots, x_m \in E$ and distinct $\mu_1, \cdots, \mu_m \in D$.

THEOREM 2. *If* φ *possesses the interpolation property then each* $p_m(\cdot; \mu_1, \cdots, \mu_m)$ *has* m *distinct zeros in the interval* E.

The stage is now set for the introduction of our "zero-mapping" transformations. Let us assume that the conditions of Theorem 1 hold, that the p_m's are known explicitly and that φ possesses the interpolation property.

For every distinct $\mu_1, \cdots, \mu_m \in D$ we have a unique $p_m(\cdot; \mu_1, \cdots, \mu_m)$. This defines a mapping from distinct m-tuples in D^m into mth-degree monic polynomials. Let $q_m(x) := \prod_{k=1}^{m} (x - \mu_k)$. Then, by the same token, we have a mapping

$$Tq_m = p_m(\cdot; \mu_1, \cdots, \mu_m) \tag{2.4}$$

of monic mth-degree polynomials into themselves, with the property that each polynomial with distinct zeros in D is mapped into a polynomial with distinct zeros in E.

Moreover, let us assume that $p_m(\cdot; \mu_1, \cdots, \mu_m)$ remain well defined even if some μ_k's coalesce and that p_m does not vanish at the endpoints of E. Then Theorem 3 follows at once.

THEOREM 3. *The mapping T maps polynomials with all zeros in D into polynomials with all zeros in E. Moreover, polynomials with distinct zeros are mapped into polynomials with distinct zeros.*

Thus, our method of constructing transformations consists of the following set stages:

(a) Choose $d\varphi(x, \mu) = \omega(x, \mu)\, d\alpha(x)$. Verify the existence and uniqueness conditions of Theorem 1 (i.e., *regularity* in the terminology of Iserles and Nørsett [1988]).

(b) Verify the satisfaction of the interpolation property.

(c) Derive bi-orthogonal polynomials $p_m(x; \mu_1, \cdots, \mu_m)$ in an explicit form.

(d) Verify that the polynomials remain well defined for all $\mu_1, \cdots, \mu_m \in D$, regardless of distinctness, and that they do not vanish at the endpoints of E.

In practice, an extra stage is sometimes useful:

(e) Bring the transformation T, by standard algebraic manipulation, into a "nicer" form.

Of the first four stages, typically only (b) and (c) present interesting problems (nothing general can be said, of course, about stage (e)). To verify the interpolation property we exploit its connection with total positivity [Iserles and Nørsett, 1988] and standard results on totally positive systems, cf. [Karlin and Studden, 1966]. A formula for an explicit form of the bi-orthogonal polynomials that correspond to a fairly general set of distributions will be presented later in this section. First, however, we provide an example that elucidates some familiar results.

Let ψ be any distribution in $D = (0, \infty)$ and set $\varphi(x, \mu) := \psi(x/\mu)$, $x \in E = (0, \infty)$, $\mu \in D$. Then [Iserles and Nørsett, 1988] φ is regular and

$$p_m(x; \mu_1, \cdots, \mu_m) = c_m \sum_{k=0}^{m} \frac{q_k}{c_k} x^k,$$

where $\sum_{k=0}^{m} q_k x^k = \prod_{k=1}^{m} (x - \mu_k)$ and $c_k := \int_0^\infty x^k\, d\psi(x)$, $k \geqq 0$, are the moments of ψ. Thus (dividing by $c_m > 0$),

$$T\left\{\sum_{k=0}^{m} q_k x^k\right\} = \sum_{k=0}^{m} \frac{q_k}{c_k} x^k,$$

a Pólya–Schur transformation with the multiplier sequence $\{1/c_k\}_{k=0}^{\infty}$.

Let $\psi(x) = -e^{-x}$. Then $d\varphi(x, \mu) = (1/\mu) \exp(-x/\mu)\, dx$; hence $\omega(x, \mu) = \exp(-x/\mu)$ and the interpolation property follows [Karlin and Studden, 1966]. This produces the transformation

$$T\left\{\sum_{k=0}^{m} q_k x^k\right\} = \sum_{k=0}^{m} \frac{1}{k!} q_k x^k \tag{2.5}$$

that maps positive zeros into positive zeros (note that the Pólya–Schur result refers to real zeros being mapped into real zeros, but it is trivial that if the zeros of $q(x)$ are,

in addition, positive, then so are the zeros of Tq, where T has been defined in (1.1) and $\{\alpha_k\}$ satisfies the Pólya-Schur conditions).

Another choice of ψ is the Stieltjes-Wigert distribution [Chihara, 1978]:

$$d\varphi(x,\mu)=\frac{\sigma}{\mu\sqrt{\pi}}e^{-\sigma^2(\log(x/\mu))^2}\,dx,$$

where $\sigma>0$. Let $\alpha=\exp(-1/(4\sigma^2))\in(0,1)$. Then

$$c_k=\alpha^{-(k+1)^2},\qquad k\geqq 0.$$

The interpolation property is valid, since

$$\omega(x,\mu)=\frac{\sigma}{\mu\sqrt{\pi}}e^{-\sigma^2((\log x)^2+(\log\mu)^2)}x^{2\sigma^2\log\mu}$$

and $\tilde{\omega}(x,\mu):=x^{2\sigma^2\log\mu}$ satisfies the interpolation property.

Standard manipulation produces the transformation

$$T\left\{\sum_{k=0}^{m}q_kx^k\right\}=\sum_{k=0}^{m}\alpha^{k^2}q_kx^k, \tag{2.6}$$

that maps positive zeros into positive zeros.

Finally, let $\psi(x)=\int_0^x\tau^\beta(1-\tau)^n\,d\tau$, $0\leqq x\leqq 1$; $\psi(x)\equiv n!/(\beta+1)_{n+1}$, $1\leqq x$, where $\beta>-1$ and n is a nonnegative integer. It is proved in [Iserles and Nørsett, 1988] that, although the interpolation property is invalid, the bi-orthogonal polynomial $p_m(\cdot;\mu_1,\cdots,\mu_m)$ has m distinct zeros in $(0,\infty)$. Hence our analysis holds and the transformation

$$T\left\{\sum_{k=0}^{m}q_kx^k\right\}=\sum_{k=0}^{m}\frac{(\beta+n+2)_k}{(\beta+1)_k}q_kx^k \tag{2.7}$$

maps positive zeros into positive zeros. Note that

$$\alpha_k=\frac{(\beta+n+2)_k}{(\beta+1)_k},\qquad k\geqq 0,$$

implies that

$$f(z):=\sum_{k=0}^{\infty}\frac{1}{k!}\alpha_kz^k={}_1F_1\left[\begin{matrix}\beta+n+2;\\ \beta+1;\end{matrix}z\right]$$

(see [Rainville, 1967] or [Slater, 1966] for exposition of hypergeometric functions). Thus by the first Kummer identity

$$f(z)=e^z\,{}_1F_1\left[\begin{matrix}-n-1;\\ \beta+1;\end{matrix}-z\right]=\frac{(n+1)!}{(1+\beta)_{n+1}}L_{n+1}^{(\beta)}(-z)\,e^z,$$

where $L_m^{(\alpha)}$ is a Laguerre polynomial [Szegö, 1975]. Thus, f is of the form (1.2) and $\{\alpha_k\}$ is a multiplier sequence.

Mappings (2.5) and (2.6) are well known [Pólya and Schur, 1914], [Pólya and Szegö, 1976],whereas the mapping (2.7) can be equivalently derived by techniques based on the Gauss-Lucas theorem [Marden, 1966].

We now present an explicit form of bi-orthogonal polynomials for a general family of distributions.

Let $\boldsymbol{\rho} = \{\rho_k\}_{k=0}^{\infty}$ be any family of monic polynomials, such that $\rho_k \in \pi_k(x)$, $0 \leqq k$. We set

$$I_k(\mu, \boldsymbol{\rho}) = \int_E \rho_k(x)\, d\varphi(x, \mu), \quad k \geqq 0, \quad \mu \in D.$$

It is easy to see that if the $I_k(\mu)$'s are replaced by $I_k(\mu, \boldsymbol{\rho})$'s in the definition of $Dm(\mu_1, \cdots, \mu_m)$ the determinant remains intact—this follows from the monicity of the ρ_k's and elementary operations on the columns of the underlying matrix. Likewise, the $I_k(\mu)$'s and x^l's can be replaced by $I_k(\mu, \boldsymbol{\rho})$'s and $\rho_l(x)$'s in (2.3). This extra generality will be useful in the next sections, mainly the choice $\rho_k(x) = (-1)^k(-x)_k$, $k \geqq 0$.

Given a real interval D, we consider two infinite families of linear functions $g_k(\mu) = \alpha_k + \beta_k\mu$, $h_k(\mu) = \gamma_k + \delta_k\mu$, $0 \leqq k$, such that $h_k(\mu) \neq 0$, $k \geqq 0$, $\mu \in D$, and $\alpha_k\delta_l - \beta_k\gamma_l \neq 0$, $0 \leqq k \leqq l$. We focus our attention on distributions with generalized moment functions of the form

$$I_k(\mu, \boldsymbol{\rho}) = \prod_{j=0}^{k-1} \frac{g_j(\mu)}{h_j(\mu)}, \qquad k \geqq 1,$$

$$I_0(\mu, \boldsymbol{\rho}) \equiv 1.$$

LEMMA 4. *If such a distribution exists then it is regular.*

Proof. Let

$$z_k^{(m)}(\mu) := \prod_{j=0}^{k-1} g_j(\mu) \prod_{j=k}^{m-2} h_j(\mu), \qquad k = 0, \cdots, m-1.$$

Note that $z_k^{(m)} \in \pi_{j-1}(\mu)$, with leading coefficient $\prod_{j=0}^{k-1} \beta_j \prod_{j=k}^{m-2} \delta_j$. It follows that

$$I_k(\mu, \boldsymbol{\rho}) = \frac{z_k^{(m)}(\mu)}{\prod_{j=0}^{m-2} h_j(\mu)}, \quad k \geqq 0, \quad \mu \in D,$$

$$D_m(\mu_1, \cdots, \mu_m) = \frac{1}{\prod_{l=1}^{m} \prod_{j=0}^{m-2} h_j(\mu_l)}\, V_m(\mu_1, \cdots, \mu_m), \qquad m \geqq 1,$$

where

$$V_m(\mu_1, \cdots, \mu_m) := \det \begin{pmatrix} z_0^{(m)}(\mu_1), & z_1^{(m)}(\mu_1), & \cdots, & z_{m-1}^{(m)}(\mu_1) \\ \vdots & \vdots & & \vdots \\ z_0^{(m)}(\mu_m), & z_1^{(m)}(\mu_m), & \cdots, & z_{m-1}^{(m)}(\mu_m) \end{pmatrix}.$$

To prove regularity, it is sufficient to demonstrate that $V_m(\mu_1, \cdots, \mu_m) \neq 0$ for all $m \geqq 1$ and distinct $\mu_1, \cdots, \mu_m \in D$. Note that the matrix of V_m can be brought into a Vandermonde form by elementary column operations; thus

$$V_m(\mu_1, \cdots, \mu_m) = C \prod_{1 \leqq k < l \leqq m} (\mu_l - \mu_k), \tag{2.8}$$

C independent of $\mu_1, \cdots, \mu_m$—a process that breaks down if the matrix is singular, but then, of course, (2.8) is true with $C = 0$.

We will now derive explicitly the constant C. Since $V_m(\mu_1, \cdots, \mu_{m-1}, \mu) \in \pi_{m-1}[\mu]$ and $V_m(\mu_1, \cdots, \mu_{m-1}, \mu_l) = 0$, $l = 1, \cdots, m-1$, it follows that

$$V_m(\mu_1, \cdots, \mu_m) = C_m(\mu_1, \cdots, \mu_{m-1}) \prod_{l=1}^{m-1} (\mu_m - \mu_l).$$

Moreover, by introspection,

$$C_m(\mu_1, \cdots, \mu_{m-1}) = \det \begin{pmatrix} z_0^{(m)}(\mu_1), & z_1^{(m)}(\mu_1), & \cdots, & z_{m-1}^{(m)}(\mu_1) \\ \vdots & \vdots & & \vdots \\ z_0^{(m)}(\mu_{m-1}), & z_1^{(m)}(\mu_{m-1}), & \cdots, & z_{m-1}^{(m)}(\mu_{m-1}) \\ \sigma_0^{(m)} & \sigma_1^{(m)}, & \cdots, & \sigma_{m-1}^{(m)} \end{pmatrix}, \tag{2.9}$$

where $\sigma_k^{(m)} := \prod_{j=0}^{k-1} \beta_j \prod_{j=k}^{m-2} \delta_j$ is the leading coefficient of $z_k^{(m)}$.

We assume first that $\delta_k \neq 0$, $k = 0, \cdots, m-2$, multiply each kth column by β_{k-1}/δ_{k-1} and subtract it from the $(k+1)$st column. Since

$$\sigma_k^{(m)} - \frac{\beta_{k-1}}{\delta_{k-1}} \sigma_{k-1}^{(m)} = 0,$$

$$z_k^{(m)}(\mu_l) - \frac{\beta_{k-1}}{\delta_{k-1}} z_{k-1}^{(m)}(\mu_l) = \frac{\alpha_{k-1}\delta_{k-1} - \beta_{k-1}\gamma_{k-1}}{\delta_{k-1}} \prod_{j=0}^{k-2} g_j(\mu_l) \prod_{j=k}^{m-2} h_j(\mu_l),$$

$$l = 1, \cdots, m-1, \qquad k = 1, \cdots, m-1;$$

zeros are inserted into the bottom row, save for the first element, and, expanding in that row,

$$C_m(\mu_1, \cdots, \mu_{m-1}) = (-1)^{m-1} \prod_{k=0}^{m-2} (\alpha_k\delta_k - \beta_k\gamma_k)/\delta_k \times \det \begin{pmatrix} \tilde{z}_0(\mu_1), & \cdots, & \tilde{z}_{m-2}(\mu_1) \\ \vdots & & \vdots \\ \tilde{z}_0(\mu_{m-1}), & \cdots, & \tilde{z}_{m-2}(\mu_{m-1}) \end{pmatrix},$$

$$\tilde{z}_k(\mu) := \prod_{j=0}^{k-1} g_j(\mu) \prod_{j=k+1}^{m-2} h_j(\mu).$$

Let $h_j^{(l)}(\mu) := h_{j+l}(\mu)$, $l, j \geqq 0$ and let $V_m^{(l)}(\mu_1, \cdots, \mu_m)$ be the determinant which is obtained by replacing h_j by $h_j^{(l)}$ in the definition of V_m, leaving the g_js intact. Hence $V_m^{(0)} \equiv V_m$ and

$$C_m(\mu_1, \cdots, \mu_{m-1}) = (-1)^{m-1} \prod_{k=0}^{m-2} \frac{\alpha_k\delta_k - \beta_k\gamma_k}{\delta_k} V_{m-1}^{(1)}(\mu_1, \cdots, \mu_{m-1}).$$

It follows that $V_m^{(0)}$ obeys the recurrence

$$V_m^{(0)}(\mu_1, \cdots, \mu_m) = (-1)^{m-1} \prod_{k=0}^{m-2} \frac{\alpha_k\delta_k - \beta_k\gamma_k}{\delta_k} \prod_{l=1}^{m-1} (\mu_m - \mu_l) V_{m-1}^{(1)}(\mu_1, \cdots, \mu_{m-1})$$

and, by induction,

$$V_m^{(0)}(\mu_1, \cdots, \mu_m) = (-1)^{ms - 1/2s(s+1)} \left\{ \prod_{l=0}^{s-1} \prod_{k=l}^{m-2} \frac{\alpha_{k-l}\delta_k - \beta_{k-l}\gamma_l}{\delta_k} \right\} \times \left\{ \prod_{l=m+1-s}^{m} \prod_{k=1}^{l-1} (\mu_l - \mu_k) \right\} V_{m-s}^{(s)}(\mu_1, \cdots, \mu_{m-s}), \qquad s = 0, \cdots, m-1.$$

In particular, setting $s = m-1$ yields

$$V_m(\mu_1, \cdots, \mu_m) = (-1)^{1/2(m-1)m} \left\{ \prod_{l=0}^{m-2} \prod_{k=l}^{m-2} \frac{\alpha_{k-l}\delta_k - \beta_{k-l}\gamma_l}{\delta_k} \right\} \prod_{1 \leqq k < l \leqq m} (\mu_l - \mu_k) \neq 0,$$

since $\alpha_n\delta_k - \beta_n\gamma_k \neq 0$ for all $n \leqq k$. This proves regularity, subject to $\delta_k \neq 0$, $k \geqq 0$.

It is obvious from the method of proof how to allow vanishing δ_k's: if $\delta_n = 0$, say, then $\sigma_k^{(m)} = 0$ for all $k \leqq n$. The elimination in (2.9) can then be carried out by using a pivot on the bottom right corner. Although the value of $C_m(\mu_1, \cdots, \mu_m)$ changes, this quantity remains nonzero and the lemma is true. □

Having proved existence and uniqueness of the underlying bi-orthogonal polynomials, it is easy to furnish an explicit formula.

LEMMA 5. *Let*

$$q_m(x) := \prod_{k=1}^{m} (x - \mu_k) = \sum_{k=0}^{m} d_k z_k^{(m+1)}(x) = \sum_{k=0}^{m} d_k \prod_{j=0}^{k-1} g_j(x) \prod_{j=k}^{m-1} h_j(x).$$

Then

$$p_m(x; \mu_1, \cdots, \mu_m) = \frac{1}{d_m} \sum_{k=0}^{m} d_k \rho_k(x).$$

Proof. With the above form of p_m,

$$H_m(\mu) = \int_E p_m(x; \mu_1, \cdots, \mu_m)\, d\varphi(x, \mu) = \frac{1}{d_m} \sum_{k=0}^{m} d_k I_k(\mu; \boldsymbol{\rho})$$

$$= \frac{1}{d_m} \left(\prod_{j=0}^{m-1} h_j^{-1}(\mu) \right) \sum_{k=0}^{m} d_k z_k^{(m+1)}(\mu) = q_m(\mu);$$

hence $H_m(\mu_l) = 0$, $l = 1, 2, \cdots, m$, and the bi-orthogonality conditions are satisfied. □

We now have the following theorem.

THEOREM 6. *Let $\varphi(x, \mu)$, $x \in E$ be a distribution for every $\mu \in D$, such that*

$$I_k(\mu, \boldsymbol{\rho}) = \prod_{j=0}^{k-1} \frac{g_j(\mu)}{h_j(\mu)}, \qquad k = 0, 1, \cdots,$$

where $\boldsymbol{\rho}$ is given, g_j, h_j, $j = 0, 1, \cdots$, are linear and g_k/h_l are bounded and monotone for all $k = 0, \cdots, l$. Then, given that φ possesses the interpolation property, the transformation

$$T\left\{ \sum_{k=0}^{m} d_k \prod_{j=0}^{k-1} g_j(x) \prod_{j=k}^{m-1} h_j(x) \right\} = \sum_{k=0}^{m} d_k \rho_k(x)$$

maps polynomials whose zeros are all in D into polynomials with zeros all in cl E.

Proof. The proof follows as a straightforward consequence of Theorem 3 and Lemmata 4 and 5 upon noting that $h_k(\mu) \neq 0$, $\mu \in D$, $\alpha_k\delta_l - \beta_k\gamma_l \neq 0$, $0 \leqq k \leqq l$, is equivalent to the boundedness and monotonicity of g_k/h_l for all $0 \leqq k \leqq l$. □

The last theorem can be typically strengthened, cl E being replaced by E, upon showing that, under the stipulated conditions, $\sum_{k=0}^{m} d_k\rho_k$ cannot vanish at the endpoints of E.

The main problem rests in the identification of distributions with the desired generalised moments. Although these moments bear certain resemblance to those connected with Hahn-type orthogonal polynomials [Andrews and Askey, 1985], a characterisation of the underlying distributions is unknown.

Fortunately, it is frequently possible to identify such distributions. For example, the distributions $d\varphi(x, \mu) = \psi(x/\mu)$, that were introduced earlier in this section, correspond to $\rho_k(x) = x^k$, $k \geqq 0$, $g_0(\mu) \equiv 1$, $g_k(\mu) = (c_k/c_{k-1})\mu$, $k \geqq 1$, $h_k(u) \equiv 1$, $k \geqq 0$, where $c_k = \int_0^\infty x^k \, d\psi(x)$, $k \geqq 0$.

In the next four sections we introduce many further distributions that satisfy the conditions of Theorem 6.

3. Transformations of hypergeometric type. In this section we present five transformations that are related to distributions that induce familiar sets of orthogonal polynomials. All these sets involve hypergeometric functions, motivating our classification. The explicit form of the bi-orthogonal polynomials has been derived in the first four cases in [Iserles and Nørsett, 1988].

Jacobi transformation. Let $d\varphi(x, \mu) = x^{\mu-1}(1-x)^\alpha \, dx$, $\alpha > -1$, $D = (0, \infty)$, $E = (0, 1)$. Thus

$$I_k(\mu) = \frac{(\mu)_k}{(\alpha+\mu+1)_k}, \qquad k \geqq 0,$$

and $g_k(\mu) = k+\mu$, $h_k(\mu) = \alpha+k+1+\mu$, $k \geqq 0$, $g_l h_k' - h_k g_l' \equiv k-l+1+\alpha \neq 0$, $0 \leqq l \leqq k$. Thus Theorem 7 follows.

THEOREM 7. *The transformation*

$$T\left\{\sum_{h=0}^{m} (x)_k(\alpha+1+k+x)_{m-k} d_k\right\} = \sum_{k=0}^{m} d_k x^k, \qquad \alpha > -1, \tag{3.1}$$

maps polynomials with positive zeros into polynomials with zeros in $(0, 1)$.

Proof. The interpolation property is valid [Iserles and Nørsett, 1988]. Thus, the conditions of Theorem 6 are satisfied and it is only necessary to verify that no zeros are possible at the endpoints. But let

$$q(x) := \sum_{k=0}^{m} (x)_k(\alpha+1+k+x)_{m-k} d_k.$$

Then $d_0 = q(0)$, $\sum_{k=0}^m d_k$ is the coefficient of x^m in q and neither quantity may vanish. □

The transformations (3.1) can be recast in a more convenient form.

LEMMA 8. *Let*

$$q(x) = \sum_{k=0}^{m} (x)_k(\alpha+1+k+x)_{m-k} d_k = \sum_{k=0}^{m} (x)_k r_k. \tag{3.2}$$

Then $d_k = (1/(\alpha+1)_m)(-1)^k \sum_{j=0}^k \binom{m-j}{k-j}(-m-\alpha)_j r_j$, $k = 0, \cdots, m$.

Proof. It is enough to show that, with the stipulated values of $d_0, \cdots, d_m$, (3.2) is true for $m+1$ different values of x. We choose $x_n = -n$, $n = 0, \cdots, m$. Then

$$q(-n) = n! \sum_{k=0}^{n} \frac{(-1)^k}{(n-k)!} r_k$$

and

$$
\begin{aligned}
n!\sum_{k=0}^{n}\frac{(-1)^k}{(n-k)!}(\alpha+1+k-n)_{m-k}d_k
&=(-1)^m\frac{n!}{(-\alpha-m)_n}\sum_{k=0}^{n}(-1)^k\frac{(-\alpha)_{n-k}}{(n-k)!}\sum_{j=0}^{k}\binom{m-j}{k-j}(-m-\alpha)_j r_j\\
&=\frac{(-1)^m n!}{(-\alpha-m)_n}\sum_{j=0}^{n}\frac{(m-j)!}{(n-j)!}(-m-\alpha)_j r_j\sum_{k=j}^{n}(-1)^k\binom{n-j}{k-j}\frac{(-\alpha)_{n-k}}{(m-k)!}\\
&=\frac{(-1)^{m-n}n!}{(-\alpha-m)_n}\sum_{j=0}^{n}\frac{(m-j)!}{(n-j)!}(-m-\alpha)_j r_j\sum_{k=0}^{n-j}(-1)^k\binom{n-j}{k}\frac{(-\alpha)_k}{(m-n+k)!}\\
&=\frac{(-1)^{m-n}n!}{(-\alpha-m)_n}\sum_{j=0}^{n}\frac{(m-j)!(-m-\alpha)_j}{(n-j)!(m-n)!}r_j\,{}_2F_1\left[\begin{matrix}-n+j,-2;\\ m-n+1;\end{matrix}1\right]\\
&=\frac{(-1)^{m-n}n!}{(-\alpha-m)_n}\sum_{j=0}^{n}\frac{(m-j)!(-m-\alpha)_j}{(n-j)!(m-n)!}r_j\frac{(m-n+1+\alpha)_{n-j}}{(m-n+1)_{n-j}}\\
&=n!\sum_{j=0}^{n}\frac{(-1)^j}{(n-j)!}r_j=q(-n),
\end{aligned}
$$

where we have used the Vandermonde theorem to sum up the hypergeometric function with a unit argument. □

We now discard the factor of $((\alpha+1)_m)^{-1}$ to obtain the following theorem.

THEOREM 7A. *The transformation*

$$T\left\{\sum_{k=0}^{m}(x)_k r_k\right\}=\sum_{k=0}^{m}(-1)^k(-m-\alpha)_k x^k(1-x)^{m-k}r_k,\qquad \alpha>-1,$$

maps polynomials with positive zeros into polynomials with zeros in $(0, 1)$.

Proof. The proof follows immediately from (3.1) and Lemma 8, upon noting that (without the common factor of $((\alpha+1)_m)^{-1}$)

$$
\begin{aligned}
\sum_{k=0}^{m}d_k x^k&=\sum_{k=0}^{m}(-1)^k\sum_{j=0}^{k}\binom{m-j}{k-j}(-m-\alpha)_j r_j x^k\\
&=\sum_{j=0}^{m}(-1)^j(-m-\alpha)_j x^j r_j\sum_{k=0}^{m-j}(-1)^k\binom{m-j}{k}x^k\\
&=\sum_{j=0}^{m}(-1)^j(-m-\alpha)_j x^j(1-x)^{m-j}r_j.
\end{aligned}
$$

□

Laguerre transformation. Let $d\varphi(x,\mu)=(1/\Gamma(\mu))x^{\mu-1}e^{-x}\,dx$, $D=E=(0,\infty)$. The interpolation property holds by [Iserles and Nørsett, 1988]. Moreover,

$$I_k(\mu)=(\mu)_k,\qquad k\geqq 0;$$

hence $g_k(\mu)=k+\mu$, $h_k(\mu)\equiv 1$, $g_l'h_k-g_kh_k'\equiv 1$ for all $0\leqq l\leqq k$. Thus, by Theorem 6 and as migration of zeros to the boundary of E can be easily excluded, we obtain the following theorem.

THEOREM 9. *The transformation*

$$T\left\{\sum_{k=0}^{m}(x)_k r_k\right\}=\sum_{k=0}^{m}r_k x^k$$

maps polynomials with positive zeros into polynomials with positive zeros.

Meixner transformation. Let $d\varphi(x,\mu)=(1-\mu/\lambda)^{-\lambda}\mu^x\,d\psi(x)$, $D=(0,\lambda)$, $E=(0,\infty)$, where ψ is a step function with jumps of $(\lambda)_k/(k!\lambda^k)$ at $k=0, 1, \cdots$. Here λ is a positive constant. We set $\rho_k(x):=(-1)^k(-x)_k$, $k\geqq 0$. Hence

$$I_k(\mu,\boldsymbol{\rho})=(\lambda)_k\left(\frac{\mu}{\lambda-\mu}\right)^k,\qquad k\geqq 0,$$

and $g_k=(1+k/\lambda)\mu$, $h_k=1-\mu/\lambda$, $g_l'h_k-g_lh_k'=1+l/\lambda\neq 0$, $0\leqq l\leqq k$. The interpolation property is again valid [Iserles and Nørsett, 1988].

By referring to Theorem 6, we at once have Theorem 10 below.

THEOREM 10. *Let λ be a positive constant. Then the transformation*

$$T\left\{\sum_{k=0}^{m} d_k(\lambda)_k x^k(\lambda-x)^{m-k}\right\}=\sum_{k=0}^{m}(-1)^k(-x)_k d_k \tag{3.3}$$

maps polynomials with zeros in $(0,\lambda)$ into polynomials with positive zeros.

Proof. It only remains to verify that $p(x)=\sum_{k=0}^{m}(-1)^k(-x)_k d_k$ does not vanish at the endpoints. Let

$$q(x)=\sum_{k=0}^{m} d_k(\lambda)_k x^k(\lambda-x)^{m-k}.$$

Then $p(0)=\lambda^{-m}q(0)\neq 0$ and, for $x\gg 0$, $p(x)\simeq d_m=\lambda^{-m}q(\lambda)/(\lambda)_m\neq 0$. □

Transformation (3.3) can easily be reformulated.

THEOREM 10A. *Let λ be a positive constant. Then the transformation*

$$T\left\{\sum_{k=0}^{m} q_k x^k\right\}=\sum_{k=0}^{m}(-1)^k(-x)_k(\lambda+x)_{m-k}\lambda^k q_k$$

maps polynomials with zeros in $(0,\lambda)$ into polynomials with positive zeros.

Proof. Given constants $q_0,\cdots,q_m$, set

$$d_k=\frac{1}{(\lambda)_k}\sum_{j=0}^{k}\binom{m-j}{k-j}\lambda^k q_j,\qquad k=0,\cdots,m.$$

Then

$$\begin{aligned}\sum_{k=0}^{m} d_k(\lambda)_k x^k(\lambda-x)^{m-k}&=\sum_{k=0}^{m}\sum_{j=0}^{k}\binom{m-j}{k-j}\lambda^k q_j x^k(\lambda-x)^{m-k}\\&=\sum_{j=0}^{m}\lambda^j q_j x^j\sum_{k=0}^{m-j}\binom{m-j}{k}x^k(\lambda-x)^{m-j-k}\\&=\lambda^m\sum_{j=0}^{m} q_j x^j\end{aligned} \tag{3.4}$$

and, by using the Vandermonde theorem [Rainville, 1967],

$$\begin{aligned}\sum_{k=0}^{m}(-1)^k(-x)_k d_k&=\sum_{k=0}^{m}(-1)^k\frac{(-x)_k}{(\lambda)_k}\sum_{j=0}^{k}\binom{m-j}{k-j}\lambda^j q_j\\&=\sum_{j=0}^{m}(-1)^j\lambda^j q_j\sum_{k=0}^{m-j}(-1)^k\binom{m-j}{k}\frac{(-x)_{k+j}}{(\lambda)_{k+j}}\\&=\sum_{j=0}^{m}(-1)^j\lambda^j q_j\frac{(-x)_j}{(\lambda)_j}\,{}_2F_1\left[\begin{matrix}-m+j,-x+j;\\ \lambda+j;\end{matrix}1\right]\\&=\sum_{j=0}^{m}(-1)^j\lambda^j q_j\frac{(-x)_j(\lambda+x)_{m-j}}{(\lambda)_j(\lambda+j)_{m-j}}\\&=\frac{1}{(\lambda)_m}\sum_{j=0}^{m}(-1)^j q_j\lambda^j(-x)_j(\lambda+x)_{m-j}.\end{aligned} \tag{3.5}$$

The theorem now follows by comparing (3.4) and (3.5) with (3.3). □

Charlier transformation. Now $d\varphi(x,\mu)=e^{-\mu}\mu^x\,d\psi(x)$, $D=E=(0,\infty)$, where ψ is a step function with jumps of $1/k!$ at $k=0,1,\cdots$. Again, it is possible to prove that all the conditions of Theorem 6 are satisfied. However, it is easier to take advantage

of the fact that the present distribution is a limiting case of the one connected with the Meixner transformation as $\lambda \to \infty$.

THEOREM 11. *The transformation*

$$T\left\{\sum_{k=0}^{m} q_k x^k\right\} = \sum_{k=0}^{m} (-1)^k q_k (-x)_k$$

maps polynomials with positive zeros into polynomials with positive zeros.

As we have already mentioned in § 1, the last result can be alternatively proved by using the convolution-orthogonality theory of Al-Salam and Ismail [1976].

Krawtchouk transformation. Let $d\varphi(x, \mu) = (1+\mu)^{-\eta}\mu^x\, d\psi(x)$, $D = E = (0, \infty)$, where $\eta \geqq 1$ and ψ is a step function with jumps of $(-1)^k(-\eta)_k/k!$ at $k = 0, 1, \cdots$. Note that, strictly speaking, φ is not a distribution—the jumps may be negative for $k > \eta$. However, as long as we confine our attention to $k \leqq \eta$ the theory of § 2 remains valid.

Let $\delta_k = (-1)^k(-x)_k$, $k = 0, 1, \cdots$. Then

$$\begin{aligned} I_k(x, \boldsymbol{\rho}) &= (-1)^k(1+\mu)^{-\eta} \sum_{l=0}^{\infty} (-1)^l(-l)_k \frac{(-\eta)_l}{l!}\mu^l \\ &= (-1)^k(-\eta)_k(1+\mu)^{-\eta}\mu^k \sum_{l=0}^{\infty} \frac{(-1)^l}{l!}(-\eta+k)_l\mu^l \\ &= (-1)^k(1+\mu)^{-\eta}\mu^k\, {}_1F_0\left[\begin{matrix}-\eta+k;\\ -;\end{matrix} -\mu\right] \\ &= (-1)^k\mu^k(1+\mu)^{-k}, \qquad k = 0, 1, \cdots. \end{aligned}$$

Thus $g_k = (\eta - k)x$, $h_k = 1 + x$, $g_l'h_k - g_lh_k' = \eta - l \neq 0$, $l = 0, \cdots, k$, $k = 0, 1, \cdots, [\eta] - 1$.

φ possesses the interpolation property—the proof follows easily from [Iserles and Nørsett, 1988]. We again use Theorem 6 to prove Theorem 12 below.

THEOREM 12. *Let $1 \leqq m \leqq \eta$. Then the transformation*

$$T\left\{\sum_{k=0}^{m} d_k(-1)^k(-\eta)_k x^k(1+x)^{m-k}\right\} = \sum_{k=0}^{m} d_k(-1)^k(-x)_k \tag{3.6}$$

maps polynomials with positive zeros into polynomials with positive zeros.

Proof. It is only necessary to demonstrate that $p(x) = \Sigma(-1)^k d_k(-x)_k$ may not vanish at the endpoints. Let $q(x) = \Sigma(-1)^k d_k(-\eta)_k x^k(1+x)^{m-k}$. Then

$$p(0) = d_0 = q(0) \neq 0, \quad p(x) \simeq d_m x^m \simeq (-1)^m q(x)/(-\eta)_m \neq 0, \quad x \gg 0$$

and the theorem is true. □

Transformation (3.6) can be recast as below.

THEOREM 12A. *Let $1 \leqq m \leqq \eta$. Then the transformation*

$$T\left\{\sum_{k=0}^{m} q_k x^k\right\} = \sum_{k=0}^{m} (-x)_k(x-\eta)_{m-k} q_k$$

maps polynomials with positive zeros into polynomials with positive zeros.

Proof. Set

$$\sum_{k=0}^{m} d_k(-1)^k(-\eta)_k x^k(1+x)^{m-k} = \sum_{k=0}^{m} q_k x^k.$$

Changing the variable into $y = x/(1+x)$ yields

$$\begin{aligned} \sum_{k=0}^{m} d_k(-1)^k(-\eta)_k y^k &= \sum_{k=0}^{m} q_k y^k(1-y)^{m-k} \\ &= \sum_{k=0}^{m} (-1)^k \left(\sum_{l=0}^{k} (-1)^l \binom{m-l}{k-l} q_l\right) y^k; \end{aligned}$$

hence

$$d_k = \frac{1}{(-\eta)_k} \sum_{l=0}^{k} (-1)^l \binom{m-l}{k-l} q_l, \qquad k = 0, \cdots, m.$$

Consequently,

$$\begin{aligned}\sum_{k=0}^{m} d_k (-1)^k (-x)_k &= \sum_{k=0}^{m} \frac{(-1)^k}{(-\eta)_k} (-x)_k \sum_{l=0}^{k} (-1)^l \binom{m-l}{k-l} q_l \\ &= \sum_{l=0}^{m} q_l \frac{(-x)_l}{(-\eta)_l} {}_2F_1\left[\begin{matrix}-m+l, -x+l; \\ -\eta+l;\end{matrix} 1\right] \\ &= \frac{1}{(-\eta)_m} \sum_{l=0}^{m} q_l (-x)_l (x-\eta)_{m-l},\end{aligned}$$

where, as before, the Vandermonde theorem has been used to sum the hypergeometric series with unit argument.

Comparison with (3.6) furnishes the proof. □

4. Transformations of q-hypergeometric type. As notation for q-hypergeometric functions varies (cf. [Andrews and Askey, 1985], [Exton, 1983], [Slater, 1966]), we wish to clarify that, throughout this paper,

$${}_p\Phi_q\left[\begin{matrix}a_1, \cdots, a_p; \\ b_1, \cdots, b_q;\end{matrix} \rho, x\right] := \sum_{k=0}^{\infty} \frac{(a_1; \rho)_k (a_2; \rho)_k \cdots (a_p; \rho)_k}{(b_1; \rho)_k (b_2; \rho)_k \cdots (b_q; \rho)_k} \frac{x^k}{(\rho; \rho)_k},$$

$$[\rho]_n := (\rho; \rho)_n, \quad n = 0, 1, \cdots,$$

$$\begin{bmatrix} m \\ n \end{bmatrix} := \frac{[\rho]_m}{[\rho]_n [\rho]_{m-n}}, \quad n = 0, \cdots, m.$$

Wall transformation. Let α be any number in $(-1, 1)$. We set

$$d\varphi(x, \mu) = \frac{(\mu; \rho)_\infty}{(\alpha\mu; \rho)_\infty} x^{\log \mu / \log \rho} \, d\psi(x), \qquad D = E = (0, 1),$$

where $\rho \in (0, 1)$ and ψ is a step function with jumps of $(\alpha; \rho)_k / [\rho]_k$ at ρ^k, $k = 0, 1, \cdots$. We note at once that the interpolation property is valid: $d\varphi(x, \mu)/d\psi(x) = \omega(x, \mu) = C(\mu) x^{\log \mu / \log \rho}$, where $C \neq 0$ is independent of x. But

$$\det \begin{pmatrix} x_1^{\nu_1}, & x_1^{\nu_2}, & \cdots, & x_1^{\nu_m} \\ \vdots & \vdots & & \vdots \\ x_m^{\nu_1}, & x_m^{\nu_2}, & \cdots, & x_m^{\nu_m} \end{pmatrix} \neq 0$$

for all distinct $x_1, \cdots, x_m > 0$ and distinct $\nu_1, \cdots \nu_m \in \mathcal{R}$, hence φ possesses the interpolation property.

The moments of φ can be evaluated by summing the ${}_1\Phi_0$ function with the Heine formula [Slater, 1966]:

$$\begin{aligned} I_k(\mu) &= \frac{(\mu; \rho)_\infty}{(\alpha\mu; \rho)_\infty} \sum_{l=0}^{\infty} \rho^{kl} \mu^l \frac{(\alpha; \rho)_l}{[\rho]_l} \\ &= \frac{(\mu; \rho)_\infty}{(\alpha\mu; \rho)_\infty} {}_1\Phi_0\left[\begin{matrix}\alpha; \\ -;\end{matrix} \rho, \mu\rho^k\right] = \frac{(\mu; \rho)_k}{(\alpha\mu; \rho)_k}, \qquad k = 0, 1, \cdots. \end{aligned}$$

Hence $g_k = 1 - \rho^k x$, $h_k = 1 - \alpha\rho^k x$, $g_l' h_k - g_l h_k' = \alpha\rho^k - \rho^l \neq 0$ (since $\alpha \in (-1, 1)$, $\rho \in (0, 1)$) for all $l = 0, \cdots, k$, $k = 0, 1, \cdots$. We are within the conditions of Theorem 6, and consequently, Theorem 13 follows.

THEOREM 13. *Let* $\alpha \in (-1, 1)$, $\rho \in (0, 1)$. *The transformation*

$$T\left\{\sum_{k=0}^{m} d_k(x;\rho)_k(\alpha x\rho^k;\rho)_{m-k}\right\} = \sum_{k=0}^{m} d_k x^k \tag{4.1}$$

maps polynomials with zeros in $(0, 1)$ *into polynomials with zeros in* $(0, 1)$.

Proof. Again, it is enough to demonstrate that $p(x) = \sum d_k x^k$ does not vanish at the endpoints. But

$$p(0) = \frac{1}{(\alpha;\rho)_m} q(1), \qquad p(1) = q(0),$$

where $q(x) = \sum d_k(x;\rho)_k(\alpha x\rho^k;\rho)_{m-k}$ and the theorem follows. □

As in § 3, the main effort is in recasting the transformation (4.1) into an equivalent form.

LEMMA 14. *Let*

$$q(x) = \sum_{k=0}^{m} d_k(x;\rho)_k(\alpha x\rho^k;\rho)_{m-k} = \sum_{k=0}^{m} s_k(x;\rho)_k.$$

Then

$$d_k = \sum_{l=0}^{k} (-1)^{k-l} \begin{bmatrix} m-l \\ k-l \end{bmatrix} \rho^{1/2(k-l)(k-l-1)} \frac{\alpha^{k-l} s_l}{(\alpha;\rho)_{m-l}}, \qquad k = 0, 1, \cdots, m.$$

Proof. With the above values of $d_0, \cdots, d_m$,

$$\begin{aligned}
q(\rho^{-n}) &= (-1)^n [\rho]_n (\alpha;\rho)_{m-n} \sum_{l=0}^{n} (-1)^l \frac{[\rho]_{m-l}\alpha^{n-l}s_l}{(\alpha;\rho)_{m-l}} \rho^{-1/2((n+1)n-(l+1)l)} \\
&\quad \times \sum_{k=l}^{n} (-1)^k \frac{(\rho/\alpha;\rho)_{n-k}}{[\rho]_{k-l}[\rho]_{m-l}[\rho]_{n-k}} \rho^{1/2(k^2-k-2l)} \\
&= [\rho]_n (\alpha;\rho)_{m-n} \sum_{l=0}^{n} (-1)^l \frac{[\rho]_{m-l}}{[\rho]_{n-l}} \frac{\alpha^{n-l}s_l\rho^{-n+(1/2)l(l+1)-nl}}{[\rho]_{m-n}(\alpha;\rho)_{m-l}} \\
&\quad \times \sum_{k=0}^{n-l} (-1)^k \begin{bmatrix} n-l \\ k \end{bmatrix} \frac{(\rho/\alpha;\rho)_k}{(\rho^{m-n+1};\rho)_k} \rho^{(1/2)k(k+1)-(n-l)} \\
&= [\rho]_n (\alpha;\rho)_{m-n} \sum_{l=0}^{n} (-1)^l \begin{bmatrix} m-l \\ n-l \end{bmatrix} \frac{\alpha^{n-l}s_l\rho^{-n+(1/2)l(l+1)-nl}}{(\alpha;\rho)_{m-l}} \\
&\quad \times {}_2\Phi_1\left[\begin{matrix} \rho^{-n+l}, \rho/\alpha; \\ \rho^{-m+n+1}; \end{matrix} \rho, \rho\right] \\
&= [\rho]_n (\alpha;\rho)_{m-n} \sum_{l=0}^{n} (-1)^l \begin{bmatrix} m-l \\ n-l \end{bmatrix} \frac{\alpha^{n-l}s_l\rho^{-n+(1/2)l(l+1)-nl}}{(\alpha;\rho)_{m-l}} \\
&\quad \times \frac{(\alpha\rho^{m-n};\rho)_{n-l}}{(\rho^{m-n+1};\rho)_{n-l}} \rho^{n-l}\alpha^{-n+l} \\
&= \sum_{l=0}^{n} (-1)^l \frac{[\rho]_n}{[\rho]_{n-l}} \rho^{(1/2)l(l-1)-nl} s_l \\
&= \sum_{l=0}^{m} s_l(\rho^{-n};\rho)_l, \qquad n = 0, 1, \cdots, m.
\end{aligned}$$

Note the use of the q-analogue of the Vandermonde theorem [Slater, 1966], [Exton, 1983] to sum up the ${}_2\Phi_1$ function with argument ρ.

Since two mth degree polynomials that coincide at $m+1$ distinct points must be identical, the lemma follows. □

THEOREM 13A. *Let $\alpha \in (-1, 1)$, $\rho \in (0, 1)$. The transformation*

$$T\left\{\sum_{k=0}^{m} s_k(x; \rho)_k\right\} = \sum_{k=0}^{m} s_k(\alpha\rho^{m-k}; \rho)_k x^k (\alpha x; \rho)_{m-k}$$

maps polynomials with zeros in $(0, 1)$ into polynomials with zeros in $(0, 1)$.

Proof. We substitute $d_0, \cdots, d_m$ from the last lemma into (4.1) and use the Heine theorem [Slater, 1966]:

$$\begin{aligned}(\alpha; \rho)_m \sum_{k=0}^{m} d_k x^k &= \sum_{l=0}^{m} (\alpha\rho^{m-l}; \rho)_l s_l \sum_{k=l}^{m} (-1)^{k-l} \begin{bmatrix} m-l \\ k-l \end{bmatrix} \alpha^{k-l} x^k \rho^{1/2(k-l)(k-l-1)} \\ &= \sum_{l=0}^{m} (\alpha\rho^{m-l}; \rho)_l s_l x^l \, {}_1\Phi_0\left[\begin{matrix} \rho^{-m+l}; \\ -; \end{matrix} \rho, \rho^{m-l}\alpha x\right] \\ &= \sum_{l=0}^{m} (\alpha\rho^{m-l}; \rho)_l s_l x^l \frac{(\alpha x; \rho)_\infty}{(\alpha x \rho^{m-l}, \rho)_\infty} \\ &= \sum_{l=0}^{m} s_l (\alpha\rho^{m-l}; \rho)_l x^l (\alpha x; \rho)_{m-l}\end{aligned}$$

and the theorem follows. □

Llaw transformation. The present transformation is, in a sense, a "reverse" Wall transformation—hence the name. Let $\rho > 1$ and

$$f(z) := {}_0\Phi_0\left[\begin{matrix} -; \\ -; \end{matrix} \rho, z\right].$$

f is an entire function and

$$f\left(\frac{z}{\rho}\right) - f(z) = \sum_{k=1}^{\infty} \frac{\rho^{-k} z^k}{[\rho]_{k-1}} = \rho^{-1} z f\left(\frac{z}{\rho}\right).$$

Hence $f(z) = (1 - z/\rho) f(z/\rho)$ and, by induction,

$$f(z) = \left(\frac{z}{\rho}; \rho^{-1}\right)_s f(\rho^{-s} z), \qquad s \geqq 0.$$

Let $s \to \infty$. Since $\rho > 1$, $\rho^{-s} z \to 0$, thus $f(0) = 1$ implies that

$$f(z) = \left(\frac{z}{\rho}; \rho^{-1}\right)_\infty. \tag{4.2}$$

Now let

$$d\varphi(x, \mu) = \frac{1}{(-\rho^{-1}\mu; \rho^{-1})_\infty} x^{\log \mu / \log \rho} \, d\psi(x), \quad D = (0, \infty), \quad E = (1, \infty),$$

where ψ is a step function with the positive jumps of $(-1)^k/[\rho]_k$ at ρ^k, $k = 0, 1, \cdots$ We have by (4.2)

$$I_k(\mu) = \frac{1}{(-\rho^{-1}\mu; \rho^{-1})_\infty} {}_0\Phi_0\left[\begin{matrix} -; \\ -; \end{matrix} \rho, -\mu\rho^k\right] = (-\mu; \rho)_k, \qquad k = 0, 1, \cdots.$$

Thus $g_k = 1 + \rho^k x$, $h_k \equiv 1$, $g_l' h_k - g_l h_k' = \rho^l \neq 0$, $l = 0, \cdots, k$, $k = 0, 1, \cdots$.

THEOREM 15. *Let $\rho > 1$. The transformation*

$$T\left\{\sum_{k=0}^{m} s_k(-x;\rho)_k\right\} = \sum_{k=0}^{m} s_k x^k$$

maps polynomials with positive zeros into polynomials with zeros in $(1,\infty)$.

Proof. As the interpolation property is valid by an argument identical to that for the Wall transformation, we are within the conditions of Theorem 6. The polynomial $p(x)=\sum s_k x^k$ cannot vanish at 1 or ∞:

$$p(1)=\sum s_k = q(0),$$
$$p(x)\simeq s_m x^m \simeq \rho^{-1/2(m-1)m} q(x), \qquad x \gg 0,$$

where $q(x)=\sum s_k(-x;\rho)_u$. Hence zeros of p cannot migrate to the endpoints of E and the proof is complete. □

q-Krawtchouk transformation. Let *integer* $N \geqq 1$ be given. Like the Krawtchouk transformation, the q-analogue is valid only for a limited range of m, namely $m \leqq N$. We set

$$d\varphi(x,\mu) = \frac{(-\mu\rho^N;\rho)_\infty}{(-\mu;\rho)_\infty} x^{\log\mu/\log\rho}\, d\psi(x), \quad \rho\in(0,1), \quad D=(0,\infty), \quad E=(0,1),$$

where ψ is a step function with jumps of $(-1)^k\rho^{kN}(\rho^{-N};\rho)_k/[\rho]_k$ at ρ^k, $k=0,1,\cdots N$. Note that the jumps are positive for all $k=0,\cdots,N$. The interpolation property is again valid. The moments are obtained by exploiting the Heine formula [Slater, 1966]:

$$I_k(\mu) = \frac{(-\mu\rho^N;\rho)_\infty}{(-\mu;\rho)_\infty}\, {}_1\Phi_0\left[\begin{matrix}\rho^{-N};\\ -;\end{matrix}\rho, -\mu\rho^{N+k}\right]$$
$$= \frac{(-\mu\rho^N;\rho)_k}{(-\mu;\rho)_k}, \qquad k=0,1,\cdots.$$

Therefore $g_k = 1+\rho^{N+k}x$, $h_k = 1+\rho^k x$, $g_l'h_k - g_l h_k' = \rho^{N+k}-\rho^l \neq 0$, $l=0,\cdots,k$, $k=0,1,\cdots,N$, and all the conditions of Theorem 6 hold.

THEOREM 16. *Let $\rho\in(0,1)$. The transformation*

$$T\left\{\sum_{k=0}^{m} d_k(-x\rho^N;\rho)_k(-\rho^k x;\rho)_{m-k}\right\} = \sum_{k=0}^{m} d_k x^k, \qquad m\leqq N \tag{4.3}$$

maps polynomials with positive zeros into polynomials with zeros in $(0,1)$.

Proof. The proof follows by demonstrating, similarly to past proofs, that zeros cannot migrate to the endpoints of E. □

To reformulate (4.3), we note the similarity with (4.1). By setting $y := -x\rho^N$, $\alpha=\rho^{-N}$ we obtain

$$\sum_{k=0}^{m} d_k(y;\rho)_k(\alpha y\rho^k;\rho)_{m-k} = \sum_{k=0}^{m} s_k(y;\rho)_k,$$

where $d_0,\cdots,d_m$ are given in Lemma 14. An immediate consequence of the method of proof of Theorem 13A is Theorem 16A below.

THEOREM 16A. *Let $\rho\in(0,1)$. The transformation*

$$T\left\{\sum_{k=0}^{m} s_k(-x\rho^N;\rho)_k\right\} = \sum_{k=0}^{m} s_k(\rho^{m-N-k};\rho)_k x^k(\rho^{-N}x;\rho)_{m-k}$$

maps polynomials with positive zeros into polynomials with zeros in $(0,1)$.

Wigert transformation. The distribution

$$d\varphi(x,\mu)=\frac{\rho^{1/2(1+\mu)^2}}{\sqrt{-2\pi\log\rho}}(-\alpha\rho^{\mu};\rho)_\infty\left(\frac{\alpha}{x}\rho^{-1/2};\rho\right)_\infty x^\mu e^{1/2((\log x)^2/\log\rho)}\,dx,$$

$u\in\mathcal{R}$, $E=(0,\infty)$, $\rho\in(0,1)$, $\alpha<1$, features (with $\mu=0$) in the definition of Stieltjes-Wigert polynomials [Chihara, 1978].

Let

$$G(\sigma):=\int_0^\infty x^\sigma e^{1/2((\log x)^2/\log\rho)}\,dx.$$

Changing the variable $t=\log x$ yields

$$\begin{aligned}G(\sigma)&=\int_{-\infty}^{\infty}\exp\left((\sigma+1)t+\frac{1}{2}\frac{t^2}{\log\rho}\right)dt\\&=\sqrt{-2\log\rho}\,\rho^{-1/2(\sigma+1)^2}\int_{-\infty}^{\infty}e^{-t^2}\,dt\\&=\sqrt{-2\pi\log\rho}\,\rho^{-1/2(\sigma+1)^2}.\end{aligned}$$

Next, let

$$F(\nu):=\int_0^\infty x^\nu\left(\frac{\alpha}{x}g^{-1/2};\rho\right)_\infty e^{1/2((\log x)^2/\log\rho)}\,dx,\qquad \nu\geqq 0.$$

Since

$$\sum_{n=0}^{\infty}(-1)^n\frac{\rho^{(1/2)n(n-2)}}{[\rho]_n}z^n=(\rho^{-1/2}z;\rho)_\infty,\qquad z\in\mathcal{C}$$

[Slater, 1966], it follows that

$$\begin{aligned}F(\nu)&=\sum_{n=0}^{\infty}(-1)^n\frac{\rho^{(1/2)n(n-2)}\alpha^n}{[\rho]_n}G(\nu-n)\\&=\rho^{-1/2(\nu+1)^2}\sqrt{-2\pi\log\rho}\sum_{n=0}^{\infty}\frac{(-\rho^\nu\alpha)^n}{[\rho]_n}\\&=\rho^{-1/2(\nu+1)^2}\sqrt{-2\pi\log\rho}\;{}_0\Phi_0\left[\begin{matrix}-\\-\end{matrix};\rho,-\rho^\nu\alpha\right]\\&=\rho^{-1/2(\nu+1)^2}\sqrt{-2\pi\log\rho}\,\frac{1}{(-\rho^\nu\alpha;\rho)_\infty}.\end{aligned}$$

Thus,

$$\begin{aligned}I_k(\mu)&=\rho^{1/2(1+\mu)^2}\frac{(-\alpha\rho^\mu;\rho)_\infty}{\sqrt{-2\pi\log\rho}}F(\mu+k)=(-\rho^\mu\alpha;\rho)_k\rho^{-k\mu-(1/2)k(k+2)}\\&=(-\eta\alpha;\rho)_k\eta^{-k}\rho^{-(1/2)k(k+2)},\qquad k=0,1,\cdots,\end{aligned}$$

where $\eta:=\rho^\mu$. Henceforth we will use η, rather than μ—thus, $D=(0,\infty)$. As a function of η, I_k is of the form $\prod_{j=0}^{k-1}(g_j(\eta)/h_j(\eta))$, where $g_k=1+\alpha\rho^k\eta$, $h_k=\rho^{k+3/2}\eta$, $g_l'h_k-g_lh_k'=-\rho^{k+3/2}(1-\alpha\rho^l)\neq 0$, $l=0,\cdots,k$, $k=0,1,\cdots$ (since $\alpha<1$, $\rho\in(0,1)$).

The interpolation property holds, since

$$d\phi(x,\mu)=C_1(x)C_2(\mu)x^\mu\,dx;$$

all that matters for the interpolation property is x^{μ}. Thus, all the conditions of Theorem 6 are satisfied and we have Theorem 17.

THEOREM 17. *Let $\alpha < 1$, $\rho \in (0, 1)$. The transformations*

$$T\left\{\sum_{k=0}^{m} d_k x^{m-k}(-\alpha x;\rho)_k \rho^{1/2(m-k)(m+k-2)}\right\} = \sum_{k=0}^{m} d_k x^k, \tag{4.4}$$

$$T\left\{\sum_{k=0}^{m} t_k x^k(-\alpha\rho^{-m+1}x;\rho)_{m-k}\rho^{-(1/2)k^2}\right\} = \sum_{k=0}^{m} t_k x^k \tag{4.5}$$

map polynomials with positive zeros into polynomials with positive zeros.

Proof. Equation (4.4) is a straightforward consequence of Theorem 6—again it is easy to verify that zeros do not migrate to the endpoints. Equation (4.5) follows from (4.4) by reversing the summation and replacing x by $\rho^{m-1}x$. □

5. Transformations of gamma type. The theme common to all four transformations in this section is the appearance of the gamma function in the distribution φ.

Γ_1 *transformation.* Let $\alpha \in (0, 1)$ and

$$d\varphi(x, \mu) = (1-\alpha)^{\mu}\frac{\Gamma(x+\mu)}{\Gamma(\mu)}\,d\psi(x), \qquad D = E = (0, \infty),$$

where the step function ψ has jumps of $\alpha^k/k!$ at $k = 0, 1, \cdots$.

LEMMA 18. *The distribution φ possesses the interpolation property.*

Proof. It is enough to show that $\Gamma(x+\mu)$ is strictly totally positive [Karlin and Studden, 1966]:

$$d\varphi(x, \mu) = C(\mu)\Gamma(x+\mu)\,d\psi(x),$$

where C and ψ make no difference to interpolation property.

Let $f(x, t)$, $g(t, y)$, $x \in (a_1, b_1)$, $t \in (c, d)$, $y \in (a_2, b_2)$, be two strictly totally positive functions and let $\phi(t)$, $t \in (c, d)$, be a distribution. Then also

$$h(x, y) := \int_c^d f(x, t)g(t, y)\,d\phi(t), \quad x \in (a_1, b_1), \quad y \in (a_2, b_2) \tag{5.1}$$

is strictly totally positive [Pólya and Szegö, 1976].

We set $a_1 = a_2 = c = 0$, $b_1 = b_2 = d = \infty$, $f(x, t) = t^x$, $g(t, y) = t^y$, $\phi(t) = (1/t)\,e^{-t}\,dt$. Thus,

$$h(x, y) = \int_0^{\infty} e^{-t}t^{x+y-1}\,dt = \Gamma(x+y)$$

is strictly totally positive and, consequently, φ possesses the interpolation property.

Let $\rho_k = (-1)^k(-x)_k$, $k = 0, 1, \cdots$. Then

$$I_k(\mu, \boldsymbol{\rho}) = (-1)^k(1-\alpha)^{\mu}\sum_{l=0}^{\infty}\frac{1}{l!}(-l)_k(\mu)_l\alpha^l$$

$$= (1-\alpha)^{\mu}(\mu)_k\alpha^k\,{}_1F_0\left[\begin{matrix}\mu+k;\\ -;\end{matrix}\ \alpha\right] = \left(\frac{\alpha}{1-\alpha}\right)^k(\mu)_k = \lambda^k(\mu)_k, \qquad k = 0, 1, \cdots,$$

where $\lambda := \alpha/(1-\alpha) > 0$. Thus, $g_k = \lambda(k+x)$, $h_k \equiv 1$, $g_l'h_k - g_lh_k' \equiv \lambda > 0$, $l = 0, \cdots, k$, $k = 0, 1, \cdots$. We are within the conditions of Theorem 6; consequently Theorem 19 follows.

THEOREM 19. *Let $\lambda > 0$. The transformation*

$$T\left\{\sum_{k=0}^{m} r_k(x)_k\right\} = \sum_{k=0}^{m} (-1)^k(-x)_k\lambda^{m-k}r_k \tag{5.2}$$

maps polynomials with positive zeros into polynomials with positive zeros.

Proof. The statement of the theorem is true for

$$T\left\{\sum_{k=0}^{m} d_k\lambda^k(x)_k\right\} = \sum_{k=0}^{m} (-1)^k d_k(-x)_k;$$

it is trivial that zeros cannot migrate to the endpoints of E. Form (5.2) follows when we set $r_k = \lambda^k d_k$, $k = 0, \cdots, m$. □

Γ_2 *transformation.* Let $a, b > 0$ and

$$d\varphi(x, \mu) = \frac{\Gamma(\mu - a)\Gamma(\mu - b)}{\Gamma(\mu - a - b)\Gamma(\mu + x)}\, d\psi(x), \quad D = (a + b, \infty), \quad E = (0, \infty),$$

where ψ is a step function with jumps of $(a)_k(b)_k/k!$ at $k = 0, 1, \cdots$. We assume for the time being that $\mu - a - b$ is not an integer—this assumption will be lifted later.

LEMMA 20. *φ possesses the interpolation property.*

Proof. It is enough to demonstrate that the beta function $B(x, \mu)$ is strictly totally positive:

$$d\varphi(x, \mu) = \frac{\Gamma(\mu - a)\Gamma(\mu - b)}{\Gamma(\mu - a - b)\Gamma(\mu)} \frac{1}{\Gamma(x)}\, d\psi(x) = C_1(\mu)C_2(x)B(x, \mu)\, d\psi(x).$$

This follows easily from (5.1) by setting $f(x, t) = t^{x-1}$, $g(t, y) = (1 - t)^{y-1}$, $\phi(t) = t$, $a_1 = 0$, $b_1 = \infty$, $a_2 = a + b$, $b_2 = \infty$, $c = 0$, $d = 1$, since

$$B(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1}\, dt$$

[Rainville, 1967]. □

Next we evaluate $I_k(\mu, \boldsymbol{\rho})$, where $\rho_k = (-1)^k(-x)_k$, $k = 0, 1, \cdots$. To this purpose we use a standard formula [Rainville, 1967] to sum up ${}_2F_1$ with the unit argument

$$\begin{aligned} I_k(\mu, \boldsymbol{\rho}) &= (-1)^k \frac{\Gamma(\mu - a)\Gamma(\mu - b)}{\Gamma(\mu - a - b)\Gamma(\mu)} \sum_{l=0}^{\infty} \frac{(-l)_k(a)_l(b)_l}{l!(\mu)_l} \\ &= \frac{\Gamma(\mu - a)\Gamma(\mu - b)}{\Gamma(\mu - a - b)\Gamma(\mu + k)} (a)_k(b)_k\, {}_2F_1\left[\begin{matrix} a+k, b+k; \\ \mu+k; \end{matrix} 1\right] \\ &= \frac{(-1)^k(a)_k(b)_k}{(-\mu + a + b + 1)_k}, \qquad k = 0, 1, \cdots. \end{aligned}$$

To simplify matters, we change the parameter: $\mu' := \mu - a - b$, $D' = (0, \infty)$. Hence

$$I_k(\mu, \boldsymbol{\rho}) = (-1)^k \frac{(a)_k(b)_k}{(-\mu + 1)_k}, \qquad k = 0, 1, \cdots$$

(use of an identical symbol I_k should cause no confusion). Note that μ is not allowed, for the time being, to be an integer.

We have $g_k=(a+k)(b+k)$, $h_k=x-k-1$, $g_l'h_k-g_lh_k'=-(a+l)(b+l)\neq 0$, $l=0,\cdots,k$, $k=0,1,\cdots$.

THEOREM 21. *Let $a, b>0$. The transformation*

$$T\left\{\sum_{k=0}^{m} u_k(x-m)_k\right\}=\sum_{k=0}^{m}\frac{(-1)^k u_{m-k}}{(a)_k(b)_k}(-x)_k \tag{5.3}$$

maps polynomials with positive zeros into polynomials with positive zeros.

Proof. First, we note that μ may attain all values in $(0,\infty)$—our restriction to noninteger values is hereby lifted. This is permissible, since bi-orthogonal polynomials exist, by continuity, for all choice of distinct $\mu_1,\cdots,\mu_m>0$. Thus, all the conditions of Theorem 6 hold and the transformation

$$T\left\{\sum_{k=0}^{m} d_k(a)_k(b)_k(x-m)_{m-k}\right\}=\sum_{k=0}^{m}(-1)^k d_k(-x)_k \tag{5.4}$$

maps polynomials with positive zeros into polynomials with positive zeros: migration of zeros to the endpoints of E is excluded in what should be by now a standard way.

To obtain (5.3) we set

$$u_k := d_{m-k}(a)_{m-k}(b)_{m-k}, \qquad k=0,1,\cdots,m,$$

in (5.4). □

Γ_3 *transformation.* Let α be a positive constant and N be a natural number. We set

$$d\varphi(x,\mu)=\frac{\Gamma(x+\mu)}{(\alpha+\mu)_N\Gamma(\mu)}\,d\psi(x),$$

where the step function ψ jumps $\binom{N}{k}(\alpha)_{N-k}$ at $k=0,1,\cdots,N$. As in the case of the Krawtchouk and the q-Krawtchouk transformations, we confine our attention to $m=1,\cdots,N$.

The interpolation property holds, by a proof identical to that of Lemma 18.

Let $\rho_k := (-1)^k(-x)_k$, $k=0,1,\cdots$.

$$\begin{aligned} I_k(\mu,\boldsymbol{\rho}) &= \frac{N!}{(\alpha+\mu)_N}\sum_{l=k}^{N}\frac{(\mu)_l(\alpha)_{N-l}}{(l-k)!(N-l)!} \\ &= \frac{N!(\mu)_k(\alpha)_{N-k}}{(\alpha+\mu)_N(N-k)!}\,{}_2F_1\left[\begin{matrix}-N+k,\mu+k;\\ -\alpha-N+k+1;\end{matrix}1\right] \\ &= \frac{N!(\mu)_k(\alpha+k+\mu)_{N-k}}{(\alpha+\mu)_N(N-k)!} \\ &= \frac{N!}{(N-k)!}\frac{(\mu)_k}{(\alpha+\mu)_k}, \qquad k=0,1,\cdots,N. \end{aligned}$$

Again, all the conditions of Theorem 6 are satisfied: $g_k=(N-k)(x+k)$, $h_k=\alpha+k+x$, $g_l'h_k-g_lh_k'=\alpha(N-k)\neq 0$, $l=0,\cdots,k$, $k=0,\cdots,N-1$.

THEOREM 22. *Let $\alpha>0$, $N\in\mathscr{Z}$, $\mathscr{N}\geqq 1$. The transformation*

$$T\left\{\sum_{k=0}^{m}\frac{d_k}{(N-k)!}(x)_k(x+\alpha+k)_{m-k}\right\}=\sum_{k=0}^{m}(-1)^k d_k(-x)_k, \qquad m\leqq N, \tag{5.5}$$

maps polynomials with positive zeros into polynomials with positive zeros.

By using Lemma 8, we can recast (5.5) into a different form.

THEOREM 22A. *Let $\alpha>0$, $N\in\mathscr{Z}$, $N\geqq 1$. The transformation*

$$T\left\{\sum_{k=0}^{m} r_k(x)_k\right\}=\sum_{k=0}^{m}(-1)^k r_k(-m-\alpha+1)_k(-x)_k(x-N)_{m-k},\qquad m\leqq N,$$

maps polynomials with positive zeros into polynomials with positive zeros.

Proof. As a consequence of Lemma 8,

$$\sum_{k=0}^{m}\frac{d_k}{(N-k)!}(x)_k(x+\alpha+k)_{m-k}=\sum_{k=0}^{m} r_k(x)_k$$

holds with

$$d_k=(-1)^k\frac{(N-k)!}{(\alpha)_m}\sum_{j=0}^{k}\binom{m-j}{k-j}(-m-\alpha+1)_j r_j,\qquad k=0,\cdots,m.$$

Thus, by the Vandermonde theorem,

$$\begin{aligned}\sum_{k=0}^{m}(-1)^k d_k(-x)_k&=\frac{1}{(\alpha)_m}\sum_{j=0}^{m}(-m-\alpha+1)_j r_j\sum_{k=j}^{m}\binom{m-j}{k-j}(N-k)!\,(-x)_k\\&=\frac{1}{(\alpha)_m}\sum_{j=0}^{m}(N-j)!(-m-\alpha+1)_j r_j(-x)_j\,{}_2F_1\left[\begin{matrix}-m+j,-x+j;\\-N+j;\end{matrix}1\right]\\&=(-1)^m\frac{(N-m)!}{(\alpha)_m}\sum_{j=0}^{m}(-1)^j r_j(-m-\alpha+1)_j(-x)_j(x-N)_{m-j}\end{aligned}$$

and the proof follows. □

Γ_4 *transformation.* Let α, $\beta>0$, with β a noninteger. We set

$$d\varphi(x,\mu)=\frac{\Gamma(\beta)\Gamma(\alpha+\beta-\mu)\Gamma(x+\mu)}{\Gamma(\beta-\mu)\Gamma(\alpha+\beta)\Gamma(\mu)}\,d\psi(x),\quad D=(0,\beta),\quad E=(0,\infty),$$

where the step function ψ has jumps of $(\alpha)_k/(k!(\alpha+\beta)_k)$ at $k=0,1,\cdots$.

The interpolation property follows easily from Lemma 18, whereas the moments are obtained by using the Vandermonde theorem:

$$\begin{aligned}\rho_k&:=(-1)^k(-x)_k,\qquad k=0,1,\cdots,\\I_k(\mu,\boldsymbol{\rho})&=(-1)^k\frac{\Gamma(\beta)\Gamma(\alpha+\beta-\mu)}{\Gamma(\beta-\mu)\Gamma(\alpha+\beta)}\sum_{l=0}^{\infty}\frac{(-l)_k}{l!}\frac{(\alpha)_l(\mu)_l}{(\alpha+\beta)_l}\\&=\frac{\Gamma(\beta)\Gamma(\alpha+\beta-\mu)}{\Gamma(\beta-\mu)\Gamma(\alpha+\beta)}\frac{(\alpha)_k(\mu)_k}{(\alpha+\beta)_k}\,{}_2F_1\left[\begin{matrix}\alpha+k,\mu+k;\\\alpha+\beta+k;\end{matrix}1\right]\\&=\frac{(\alpha)_k(\mu)_k}{(\alpha+\beta)_k}\frac{\Gamma(\beta)\Gamma(\alpha+\beta-\mu)}{\Gamma(\beta-\mu)\Gamma(\alpha+\beta)}\frac{\Gamma(\alpha+\beta+k)\Gamma(\beta-\mu-k)}{\Gamma(\beta)\Gamma(\alpha+\beta-\mu)}\\&=(-1)^k\frac{(\alpha)_k(\mu)_k}{(-\beta+\mu+1)_k},\qquad k=0,1,\cdots.\end{aligned}$$

Thus $g_k=(\alpha+k)(k+x)$, $h_k=\beta-k-1-x$, $g_l'h_k-g_lh_k'=(\alpha+l)(\beta-k-1+l)\neq 0$ (since β is not an integer), $l=0,\cdots,k$, $k=0,1,\cdots$. All the conditions of Theorem 6 hold, migration of zeros to the boundary of E is prevented, and we have the following theorem.

THEOREM 23. *Let α, $\beta > 0$, β a noninteger. The transformation*

$$T\left\{\sum_{k=0}^{m} d_k(\alpha)_k(x)_k(\beta-m-x)_{m-k}\right\} = \sum_{k=0}^{m}(-1)^k d_k(-x)_k \tag{5.6}$$

maps polynomials with zeros in $(0, \beta)$ into polynomials with positive zeros.

To reformulate (5.6) we require the following lemma, which bears much similarity to Lemma 8.

LEMMA 24. *Let γ be neither zero nor a negative integer. Then*

$$\sum_{k=0}^{m} e_k(x)_k(\gamma-x)_{m-k} = \sum_{k=0}^{m} r_k(x)_k \tag{5.7}$$

implies that

$$e_k = \sum_{l=0}^{k}\binom{m-l}{k-l}\frac{r_l}{(\gamma+l)_{m-l}}, \qquad k=0,1,\cdots,m. \tag{5.8}$$

Proof. Let

$$q_1(x) := \sum_{k=0}^{m} e_k(x)_k(\gamma-x)_{m-k}, \qquad q_2(x) := \sum_{k=0}^{m} r_k(x)_k. \qquad \square$$

To prove (5.7), it is enough to demonstrate that $q_1(-n) = q_2(-n)$, $n = 0, 1, \cdots, m$, where the coefficients of q_1 and q_2 are linked via (5.8). But

$$\begin{aligned}
\frac{1}{n!}q_1(-n) &= \sum_{k=0}^{n}(-1)^k\frac{e_k}{(n-k)!}(\gamma+n)_{m-k} \\
&= \sum_{k=0}^{n}\frac{(-1)^k}{(n-k)!}(\gamma+n)_{m-k}\sum_{l=0}^{k}\binom{m-l}{k-l}\frac{r_l}{(\gamma+l)_{m-l}} \\
&= \sum_{l=0}^{n}(-1)^l\frac{(m-l)!}{(\gamma+l)_{m-l}}r_l\sum_{k=0}^{n-l}(-1)^k\frac{(\gamma+n)_{m-l-k}}{k!(n-l-k)!(m-l-k)!} \\
&= \sum_{l=0}^{n}(-1)^l\frac{(\gamma+n)_{m-l}}{(\gamma+l)_{m-l}(n-l)!}r_l\sum_{k=0}^{n-l}(-1)^k\binom{n-l}{k}\frac{(-m+l)_k}{(-\gamma-n-m+l+1)_k} \\
&= \sum_{l=0}^{n}(-1)^l\frac{(\gamma+n)_{m-l}}{(\gamma+l)_{m-l}(n-l)!}r_l\,{}_2F_1\left[\begin{matrix}-n+l,-m+l;\\ -\gamma-n-m+1;\end{matrix}1\right] \\
&= \sum_{l=0}^{n}(-1)^l\frac{r_l}{(n-l)!}\frac{(\gamma+n)_{m-l}(-\gamma-n+1)_{n-l}}{(\gamma+l)_{m-l}(-\gamma-n-m+l+1)_{n-l}}.
\end{aligned}$$

Thus, as

$$(\gamma+n)_{m-l}(-\gamma-n+1)_{n-l} = (\gamma+l)_{m-l}(-\gamma-n-m+l+1)_{n-l}, \qquad l=0,\cdots,n,$$

we have

$$q_1(-n) = n!\sum_{l=0}^{n}(-1)^l\frac{r_l}{(n-l)!} = q_2(-n), \qquad n=0,1,\cdots,m,$$

and the proof follows. $\square$

THEOREM 23A. *Let α, $\beta > 0$, β a noninteger. The transformation*

$$T\left\{\sum_{k=0}^{m} r_k(x)_k\right\} = \sum_{k=0}^{m}(-1)^k(\beta-m)_k r_k(-x)_k(\alpha+x)_{m-k}$$

maps polynomials with zeros in $(0, \beta)$ into polynomials with positive zeros.

Proof. We set $e_k := (\alpha)_k d_k$, $k = 0, \cdots, m$, $\gamma = \beta - m$ in (5.7). Note that γ is a noninteger. Thus, by (5.7), if

$$\sum_{k=0}^{m} d_k(\alpha)_k(x)_k(\beta - m - x)_{m-k} = \sum_{k=0}^{m} r_k(x)_k, \tag{5.9}$$

then $d_0, \cdots, d_m$ and $r_0, \cdots, r_m$ are linked via (5.8). Moreover,

$$\begin{aligned}
\sum_{k=0}^{m} (-1)^k d_k(-x)_k &= \sum_{k=0}^{m} (-1)^k \frac{(-x)_k}{(\alpha)_k} \sum_{l=0}^{k} \binom{m-l}{k-l} \frac{r_l}{(\beta - m + l)_{m-l}} \\
&= \sum_{l=0}^{m} (-1)^l \frac{(-x)_l r_l}{(\beta - m + l)_{m-l}(\alpha)_l} {}_2F_1\left[\begin{matrix} -m+l, -x+l; \\ \alpha+l; \end{matrix} 1\right] \\
&= \sum_{l=0}^{m} (-1)^l \frac{(-x)_l r_l}{(\beta - m + l)_{m-l}(\alpha)_l} \frac{(\alpha + x)_{m-l}}{(\alpha + l)_{m-l}} \\
&= \frac{1}{(\alpha)_m(\beta - m)_m} \sum_{l=0}^{m} (-1)^l (\beta - m)_l r_l (-x)_l (\alpha + x)_{m-l}.
\end{aligned} \tag{5.10}$$

The proof follows when we compare (5.9) and (5.10) with (5.6). □

6. Transformations of q-gamma type. There are several q-analogues of the gamma function [Exton, 1983]. In this section we consider distributions that include the function

$$\Gamma_\rho(z) := \frac{1}{(z; \rho)_\infty}, \qquad \rho, z \in C. \tag{6.1}$$

q-Γ_1 transformation. Let $\rho, \alpha \in (0, 1)$,

$$d\varphi(x, \mu) = \frac{(\alpha; \rho)_\infty (\mu; \rho)_\infty}{(\alpha\mu; \rho)_\infty (\mu x; \rho)_\infty} d\psi(x), \qquad D = E = (0, 1),$$

where ψ jumps $\alpha^k/[\rho]_k$ at ρ^k, $k = 0, 1, \cdots$. Thus,

$$\begin{aligned}
I_k(\mu) &= \frac{(\alpha; \rho)_\infty (\mu; \rho)_\infty}{(\alpha\mu; \rho)_\infty} \sum_{l=0}^{\infty} \frac{\alpha^l \rho^{kl}}{[\rho]_l (\mu\rho^l; \rho)_\infty} \\
&= \frac{(\alpha; \rho)_\infty}{(\alpha\mu; \rho)_\infty} \sum_{l=0}^{\infty} \frac{(\mu; \rho)_l}{[\rho]_l} \alpha^l \rho^{kl} = \frac{(\alpha; \rho)_\infty}{(\alpha\mu; \rho)_\infty} {}_1\Phi_0\left[\begin{matrix} \mu; \\ -; \end{matrix} \rho, \alpha\rho^k\right].
\end{aligned}$$

We sum the ${}_1\Phi_0$ series by the Heine theorem [Slater, 1966]:

$$I_k(\mu) = \frac{(\alpha; \rho)_\infty}{(\alpha\mu; \rho)_\infty} \frac{(\alpha\mu\rho^k; \rho)_\infty}{(\alpha\rho^k; \rho)_\infty} = \frac{(\alpha; \rho)_k}{(\alpha\mu; \rho)_k}, \qquad k = 0, 1, \cdots.$$

Thus, $g_k(x) \equiv 1 - \alpha\rho^k$, $h_k(x) = 1 - \alpha\rho^k x$, $g_l' h_k - g_l h_k' = \alpha\rho^k(1 - \alpha\rho^l) \neq 0$, $l = 0, \cdots, k$, $k = 0, 1, \cdots$.

LEMMA 25. *For every $\rho \in (0, 1)$ the function $\Gamma_\rho(x\mu)$ is strictly totally positive for all $0 < x, \mu < \infty$.*

Proof. Since $|\rho| < 1$, the Euler theorem [Slater, 1966] and (6.1) imply that

$$\Gamma_\rho(x\mu) = {}_0\Phi_0\left[\begin{matrix} -; \\ -; \end{matrix} \rho, x\mu\right] = \sum_{k=0}^{\infty} \frac{(\mu x)^k}{[\rho]_k}. \tag{6.2}$$

Let

$$f(x, t) = x^{\log t/\log \rho}, \quad g(t, y) = y^{\log t/\log \rho}, \quad 0 < x, y < \infty, \quad 0 < t < 1.$$

Clearly, both f and g are strictly totally positive. Moreover, let ϕ be a step function with jumps of $1/[\rho]_k$ at ρ^k, $k=0, 1, \cdots$. Thus, by (6.2),

$$\Gamma_\rho(x\mu)=\int_0^1 f(x, t)g(t, \mu)\, d\phi(t)$$

and strict total positivity follows from (5.1). □

THEOREM 26. *Let α, $\rho \in (0, 1)$. The transformation*

$$T\left\{\sum_{k=0}^{m} w_k(\alpha x\rho^k; \rho)_{m-k}\right\}=\sum_{k=0}^{m}\frac{w_k}{(\alpha; \rho)_k}x^k \tag{6.3}$$

maps polynomials with zeros in $(0, 1)$ into polynomials with zeros in $(0, 1)$.

Proof. Since all the conditions of Theorem 6 are satisfied, the transformation

$$T\left\{\sum_{k=0}^{m} d_k(\alpha; \rho)_k(\alpha\rho^k x; \rho)_{m-k}\right\}=\sum_{k=0}^{m} d_k x^k$$

maps polynomials with zeros in $(0, 1)$ into polynomials with zeros in $(0, 1)$—it is again trivial that zeros cannot migrate to the boundary.

Form (6.3) follows when we set $w_k=(\alpha; \rho)_k d_k$, $k=0, 1, \cdots, m$. □

An alternative form of (6.3) can be obtained by changing the basis of moments: Let $\rho_k(x)=\prod_{l=0}^{k-1}(x-\rho^l)=x^k(1/x; \rho)_k$, $k=0, 1, \cdots$. Then by the Heine theorem [Slater, 1966],

$$I_k(\mu; \boldsymbol{\rho})=\frac{(\mu; \rho)_\infty(\alpha; \rho)_\infty(-1)^k\alpha^k}{(\alpha\mu; \rho)_\infty(\mu\rho^k; \rho)_\infty}\,{}_1\Phi_0\left[\begin{matrix}\mu\rho^k;\\ -;\end{matrix}\rho, \alpha\right]$$

$$=(-1)^k\rho^{1/2(k-1)k}\alpha^k\frac{(\mu; \rho)_k}{(\alpha\mu; \rho)_k}, \qquad k=0, 1, \cdots.$$

Thus $g_k(x)=-\alpha\rho^k(1-\rho^k x)$, $h_k(x)=1-\alpha\rho^k x$, $h_l' g_k - h_l g_k' = -\alpha\rho^l(\alpha\rho^k-\rho^l)\neq 0$, $l=0, \cdots, k$, $k=0, 1, \cdots$, all the conditions of Theorem 6 are valid, and we have Theorem 26A.

THEOREM 26A. *Let α, $\rho \in (0, 1)$. The transformation*

$$T\left\{\sum_{k=0}^{m}(-1)^k\rho^{1/2(k-1)k}d_k(x; \rho)_k(\alpha\rho^k x; \rho)_{m-k}\right\}=\sum_{k=0}^{m} d_k\prod_{l=0}^{k-1}(x-\rho^l)$$

maps polynomials with zeros in $(0, 1)$ into polynomials with zeros in $(0, 1)$.

q-Γ_2 transformation. Let $\rho \in (0, 1)$, N be a natural number, $\alpha > \rho^{-N+1}$, and

$$d\varphi(x, \mu)=\frac{\alpha^{-N}(-\mu x; \rho)_\infty}{(-\mu\rho^N; \rho)_\infty(-(\mu/\alpha); \rho)_N}\, d\psi(x), \quad D=(0, \infty), \quad E=(0, 1),$$

where ψ jumps $(\rho^{-N}; \rho)_k(\alpha; \rho)_k\rho^k/[\rho]_k > 0$ at ρ^k, $k=0, 1, \cdots, N$.

LEMMA 27. *φ possesses the interpolation property.*

Proof. It is sufficient to prove that $(-\mu x; \rho)_\infty$ is strictly totally positive for all $x, \mu \in (0, 1)$. By (4.2),

$$(-\mu x; \rho)_\infty={}_0\Phi_0\left[\begin{matrix}-;\\ -;\end{matrix}\rho^{-1}, -\mu x\rho^{-1}\right]=\sum_{k=0}^{\infty}(-1)^k\frac{(\mu x\rho^{-1})^k}{[\rho^{-1}]_k}.$$

The proof now proceeds along the lines of the proof of Lemma 25, the distribution ϕ

being replaced by a step function with jumps of $(-1)^k/[\rho^{-1}]_k > 0$ ($[\rho^{-1}]_k = (\rho^{-1}, \rho^{-1})_k$), $k = 0, 1, \cdots$, at ρ^{-k} and the interval of integration being $(1, \infty)$. □

Let $\rho_k(x) = \prod_{l=0}^{k-1}(x-\rho^l)$, $k = 0, 1, \cdots$. Then for all $k = 0, \cdots, N$

$$I_k(\mu, \boldsymbol{\rho}) = \frac{(-1)^k \alpha^{-N} \rho^{1/2(k-1)k}}{(-\mu\rho^N; \rho)_\infty (-(\mu/\alpha); \rho)_N} \sum_{l=k}^{\infty} \frac{(-\mu\rho^l; \rho)_\infty (\rho^{-N}; \rho)_l (\alpha; \rho)_l}{[\rho]_{l-k}} \rho^l$$

$$= (-1)^k \rho^{1/2(k+1)k} \frac{\alpha^{-N}(-\mu\rho^k; \rho)_\infty (\rho^{-N}; \rho)_k (\alpha; \rho)_k}{(-\mu\rho^N; \rho)_\infty (-(\mu/\alpha); \rho)_N} {}_2\Phi_1\left[\begin{matrix}\rho^{-N+k}, \alpha\rho^k; \\ -\mu\rho^k;\end{matrix} \rho, \rho\right].$$

We sum the ${}_2\Phi_1$ series up with the Vandermonde q-analogue [Slater, 1966]:

$$I_k(\mu, \boldsymbol{\rho}) = (-1)^k \rho^{1/2(k+1)k} \frac{\alpha^{-N}(-\mu\rho^k; \rho)_\infty (\rho^{-N}; \rho)_k (\alpha; \rho)_k}{(-\mu\rho^N; \rho)_\infty (-(\mu/\alpha); \rho)_N} \frac{(-(\mu/\alpha); \rho)_{N-k} \alpha^{N-k} \rho^{k(N-k)}}{(-\mu\rho^k; \rho)_{N-k}}$$

$$= (-1)^k \alpha^{-k} \rho^{-1/2(k-1)k + Nk} \frac{(\rho^{-N}; \rho)_k (\alpha; \rho)_k}{(-(\mu/\alpha)\rho^{N-k}; \rho)_k}$$

$$= (-1)^k \frac{\rho^k}{\mu^k} \frac{(\rho^{-N}; \rho)_k (\alpha; \rho)_k}{(-(\alpha/\mu)\rho^{-N+1}; \rho)_k}, \qquad k = 0, 1, \cdots, N.$$

Hence $g_k(x) \equiv -\rho(1-\rho^{-N+k})(1-\alpha\rho^k)$, $h_k(x) = \alpha\rho^{-N+k+1} + x$, $g_l' h_k - g_l h_k' \neq 0$, $l = 0, \cdots, k$, $k = 0, 1, \cdots, N-1$. All the conditions of Theorem 6 are satisfied, leading to Theorem 28.

THEOREM 28. *Let $\rho \in (0, 1)$, $\alpha > \rho^{-N+1}$, $m = 1, 2, \cdots, N$. The transformation*

$$T\left\{\sum_{k=0}^{m} (-1)^k d_k \rho^{kN - 1/2(k-1)k} \alpha^{m-k} (\alpha; \rho)_k (\rho^{-N}; \rho)_k \left(-\frac{x}{\alpha}\rho^{N-m}; \rho\right)_{m-k}\right\}$$

$$= \sum_{k=0}^{m} d_k \prod_{l=0}^{k-1} (x-\rho^l)$$

maps polynomials with positive zeros into polynomials with zeros in $(0, 1)$.

Note that, in the last theorem, we were unable to exclude migration of zeros to the endpoints of E.

q-Γ_3 transformation. Let N be a natural number, $\rho \in (0, 1)$, and $\alpha > \rho^{-N+1}$. We set

$$d\varphi(x, \mu) = \frac{(\alpha; \rho)_N (\mu; \rho)_\infty \mu^{-N}}{((\alpha/\mu); \rho)_N (\mu x; \rho)_\infty} d\psi(x), \qquad D = E = (0, 1),$$

where ψ has jumps of $(\rho^{-N}; \rho)_k \rho^k / ([\rho]_k (\alpha; \rho)_k) > 0$ at ρ^k, $k = 0, 1, \cdots, N$. Again $\rho_k(x) := \prod_{l=0}^{k-1}(x-\rho^l)$, $k = 0, 1, \cdots$. Hence,

$$I_k(\mu, \boldsymbol{\rho}) = \frac{(\alpha; \rho)_N (\mu; \rho)_\infty \mu^{-N}}{((\alpha/\mu); \rho)_N} (-1)^k \rho^{1/2(k-1)k} \sum_{l=k}^{N} \frac{(\rho^{-N}; \rho)_l \rho^l}{[\rho]_{l-k} (\mu\rho^l; \rho)_\infty (\alpha; \rho)_l}$$

$$= \frac{(-1)^k \mu^{-N} \rho^{(1/2)k(k+1)} (\alpha; \rho)_N (\rho^{-N}; \rho)_k (\mu; \rho)_k}{((\alpha/\mu); \rho)_N (\alpha; \rho)_k} {}_2\Phi_1\left[\begin{matrix}\rho^{-N+k}, \mu\rho^k; \\ \alpha\rho^k;\end{matrix} \rho, \rho\right]$$

$$= \frac{(-1)^k \mu^{-N} \rho^{(1/2)k(k+1)} (\alpha; \rho)_N (\rho^{-N}; \rho)_k (\mu; \rho)_k}{((\alpha/\mu); \rho)_N (\alpha; \rho)_k} \times \frac{((\alpha/\mu); \rho)_{N-k} \mu^{N-k} \rho^{k(N-k)}}{(\alpha\rho^k; \rho)_{N-k}}$$

$$= \frac{(\rho^{-N}; \rho)_k (\mu; \rho)_k}{((\mu/\alpha)\rho^{-N+1}; \rho)_k} \left(\frac{\rho}{\alpha}\right)^k, \qquad k = 0, 1, \cdots, N.$$

Consequently, $g_k(x)=\rho(1-\rho^{-N+k})(1-x\rho^k)$, $h_k(x)\equiv\alpha(1-(x/\alpha)\rho^{-N+k+1})$, $g_l'h_k-g_lh_k'\neq 0$, $l=0,\cdots,k$, $k=0,\cdots,N-1$, and, since it follows from Lemma 25 that φ possesses the interpolation property, all the conditions of Theorem 6 hold. This furnishes our last transformation.

THEOREM 29. *Let* $\rho\in(0,1)$, $\alpha>\rho^{-N+1}$, $m=1,2,\cdots,N$. *The transformation*

$$T\left\{\sum_{k=0}^{m} d_k\rho^k(\rho^{-N};\rho)_k(x;\rho)_k\alpha^{m-k}\left(\frac{x}{\alpha}\rho^{-N+k+1};\rho\right)_{m-k}\right\}=\sum_{k=0}^{m} d_k\prod_{l=0}^{k-1}(x-\rho^l)$$

maps polynomials with zeros in $(0,1)$ *into polynomials with zeros in* $(0,1)$.

Note that, again, the last theorem does not prevent zeros being mapped to the endpoints of E.

Acknowledgments. The authors greatly benefited from discussing transformations and bi-orthogonality with numerous colleagues. Special mention and thanks go to Phil Davis (Brown University), Bill Gragg (Naval Postgraduate School), Mourad Ismail (University of South Florida), Ted Rivlin (IBM), Ed Saff (University of South Florida), and Gil Strang (MIT).

REFERENCES

W. A. AL-SALAM AND M. E. H. ISMAIL [1976], *Polynomials orthogonal with respect to discrete convolution*, J. Math. Anal. Appl., 55, pp. 125-139.

G. E. ANDREWS AND R. ASKEY [1985], *Classical orthogonal polynomials*, in Orthogonal Polynomials, Bar-le-Duc 1984, C. Brezinski et al., eds., Lecture Notes in Mathematics 1171, Springer-Verlag, Berlin, New York, pp. 36-62.

T. S. CHIHARA [1978], *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York.

T. CRAVEN AND G. CSORDAS [1983], *The Gauss-Lucas theorem and Jensen polynomials*, Trans. Amer. Math. Soc. 278, American Mathematical Society, Providence, R.I., pp. 415-429.

H. EXTON [1983], *q-Hypergeometric Functions and Applications*, Ellis Horwood, Chichester, U.K. Rocky Mountain J. Math., to appear.

E. HILLE [1962], *Analytic Function Theory*, Vol. II, Chelsea, New York.

A. ISERLES AND S. P. NØRSETT [1988], *On the theory of bi-orthogonal polynomials*, Trans. Amer. Math. Soc. 306, American Mathematical Society, Providence, R.I., pp. 455-474.

A. ISERLES, S. P. NØRSETT, AND E. B. SAFF [1988], *On transformations and zeros of polynomials*, Tech. Report DAMTP NA/10, University of Cambridge, Cambridge, U.K. Rocky Mountain J. Math., to appear.

A. ISERLES AND E. B. SAFF [1987], *Bi-orthogonality in rational approximation*, J. Appl. Comp. Math., 19, pp. 47-54.

——— [1989], *Zeros of expansions in orthogonal polynomials*, Math. Proc. Cambridge Philos. Soc. 105, pp. 559-573.

S. KARLIN AND W. J. STUDDEN [1966], *Tchebycheff Systems: With Applications in Analysis and Statistics*, John Wiley, New York.

M. MARDEN [1966], *Geometry of Polynomials*, American Mathematical Society, Providence, R.I.

G. PÓLYA AND J. SCHUR [1914], *Über zwei Arten von Faktorenfolgen in der Theorie der algebraischen Gleichungen*, J. Reine Angew. Math., 144, pp. 89-133.

G. PÓLYA AND G. SZEGÖ [1976], *Problems and Theorems in Analysis*, Vol. II, Springer-Verlag, Berlin, New York.

E. D. RAINVILLE [1967], *Special Functions*, Macmillan, New York.

L. J. SLATER [1966], *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, U.K.

G. SZEGÖ [1975], *Orthogonal Polynomials*, Fourth edition, American Mathematical Society, Providence, RI.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 510-522, March 1990
© 1990 Society for Industrial and Applied Mathematics
012

A SUMMATION THEOREM FOR HYPERGEOMETRIC SERIES VERY-WELL-POISED ON G_2*

ROBERT A. GUSTAFSON†

Abstract. An analogue of Bailey's ${}_6\psi_6$ summation theorem is proved for basic hypergeometric series that are very well poised on the Lie algebra G_2. As a limiting case, a new proof of the Macdonald identity associated to the affine root system of type G_2 is obtained. A summation theorem for ordinary hypergeometric series that are very well poised on G_2 is proved by Carlson's theorem.

Key words. hypergeometric series, basic hypergeometric series, summation theorem, Lie algebra G_2

AMS(MOS) subject classifications. 33A75, 33A30, 33A35

1. Introduction and statement of results. This paper is a sequel to the paper [10] entitled "The Macdonald identities for affine root systems of classical type and hypergeometric series very-well-poised on semisimple Lie algebras." We will prove an analogue of Bailey's ${}_6\psi_6$ summation theorem [5] for very-well-poised series on the complex Lie algebra G_2 [10] associated to the full co-root lattice of G_2 and to the seven-dimensional irreducible representation of G_2 of highest weight λ_1, where λ_1 is the first fundamental weight [7, App.]. A limiting case of this summation theorem is the Macdonald identity [15] associated to the affine root system of type G_2. The method is similar to that used by Andrews [1] to deduce the Jacobi triple product identity as a limit of the ${}_6\psi_6$ summation theorem and by Milne [17] in his proof of the Macdonald identities for the affine root systems of type A_l.

Before stating the main theorem, we need to define some notation. For convenience, let q be a real number, $0<q<1$. If $c\in\mathbb{C}$ and n is a positive integer, then the q-rising factorial is defined by

$$[c]_n=(1-c)(1-cq)\cdots(1-cq^{n-1}),\qquad [c]_\infty=\prod_{k=0}^{\infty}(1-cq^k).$$

For $n\in\mathbb{Z}$, define

$$[c]_n=[c]_\infty/[cq^n]_\infty.$$

We have

$$[c]_{-n}=\frac{(-1)^n q^{(n+1)n/2}}{[qc^{-1}]_n c^n}$$

for all $n\in\mathbb{Z}$.

Similarly, the ordinary rising factorial is defined by

$$(c)_n=\Gamma(c+n)/\Gamma(c)$$

for $n\in\mathbb{Z}$ and $c\in\mathbb{C}$.

In § 2 we will prove the following fundamental theorem.

THEOREM 1.1. Let $z=(z_1,z_2,z_3)\in\mathbb{C}^3$ *and* $a_i,b_i\in\mathbb{C}$ *for* $i=1,2$. Let $\varphi_i=\varepsilon_i-\frac{1}{3}(\varepsilon_1+\varepsilon_2+\varepsilon_3)$ *for* $i=1,2,3$, *where* $\varepsilon_i\in\mathbb{C}^3$ *is the standard triple with* 1 *in the ith entry*

* Received by the editors May 23, 1988; accepted for publication February 13, 1989. This work was partially supported by National Science Foundation grant INT-8713472.

† Department of Mathematics, Texas A&M University, College Station, Texas 77843 and Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India.

and zero in the other entries. Let L be the lattice in $\mathbb{C}^3$ generated by φ_1, φ_2, φ_3, i.e., $L=\sum_{i=1}^{3}\mathbb{Z}\varphi_i$. Note that $\varphi_1+\varphi_2+\varphi_3=0$. Finally, let

$$c_i=\begin{cases} a_i & \text{for } i=1,2,\\ qb_{i-2}^{-1} & \text{for } i=3,4.\end{cases} \tag{1.2}$$

Then we have

$$\begin{aligned} t=\sum_{(t_1,t_2,t_3)\in L}\Bigg\{ & q^{t_1+2t_2-3t_3}\left(\frac{1-q^{-2z_1-2t_1+z_2+t_2+z_3+t_3}}{1-q^{-2z_1+z_2+z_3}}\right)\\ & \cdot\left(\frac{1-q^{z_1+t_1-2z_2-2t_2+z_3+t_3}}{1-q^{z_1-2z_2+z_3}}\right)\left(\frac{1-q^{-z_1-t_1-z_2-t_2+2z_3+2t_3}}{1-q^{-z_1-z_2+2z_3}}\right)\\ & \cdot\left(\frac{1-q^{z_1+t_1-z_2-t_2}}{1-q^{z_1-z_2}}\right)\left(\frac{1-q^{-z_1-t_1+z_3+t_3}}{1-q^{-z_1+z_3}}\right)\left(\frac{1-q^{-z_2-t_2+z_3+t_3}}{1-q^{-z_2+z_3}}\right)\\ & \cdot\prod_{i=1}^{2}\prod_{\substack{1\le j,k\le 3\\ j\ne k}}\frac{[a_iq^{z_j-z_k}]_{t_j-t_k}}{[b_iq^{z_j-z_k}]_{t_j-t_k}}\Bigg\} \end{aligned} \tag{1.3}$$

$$\begin{aligned} =\frac{[q]_\infty^2\left[q\prod_{i=1}^{4}c_i^{-1}\right]_\infty\prod_{1\le i\le j\le 4}[q(c_ic_j)^{-1}]_\infty\prod_{1\le i<j<k\le 4}[q(c_ic_jc_k)^{-1}]_\infty}{\left[q\prod_{i=1}^{4}c_i^{-2}\right]_\infty\prod_{i=1}^{4}[qc_i^{-1}]_\infty\prod_{i=1}^{4}\prod_{\substack{1\le j,k\le 3\\ j\ne k}}[c_1^{-1}q^{1+z_j-z_k}]_\infty}\\ \cdot\prod_{i=1}^{3}\{[q^{1+3z_i-z_1-z_2-z_3}]_\infty[q^{1-3z_i+z_1+z_2+z_3}]_\infty\}\\ \cdot\prod_{\substack{1\le j,k\le 3\\ j\ne k}}[q^{1+z_j-z_k}]_\infty, \end{aligned}$$

where $|q^{-3}\prod_{i=1}^{2}(b_i/a_i)^2|<1$ for convergence and none of the denominators on both sides of (1.3) *vanish.*

Remark 1.4. We may assume that $z_1+z_2+z_3=0$ in Theorem 1.1 as this case implies the general result. With this restriction and in the notation of [10], the series on the left-hand side of (1.3) is denoted by

$${}_2\psi_2\left(G_2;\lambda_1;L;\begin{matrix}a_1, & a_2\\ b_1, & b_2\end{matrix};z;q;(1,\cdots,1)\right),$$

where L is the full co-root lattice for G_2.

COROLLARY 1.5 (Macdonald [15]). *With notation as in Theorem* 1.1, *we have*

$$\begin{aligned} \sum_{t\in L}\Bigg\{ & q^{t_1+2t_2-3t_3}\frac{(1-q^{-2z_1-2t_1+z_2+t_2+z_3+t_3})}{(1-q^{-2z_1+z_2+z_3})}\\ & \cdot\frac{(1-q^{z_1+t_1-2z_2-2t_2+z_3+t_3})(1-q^{-z_1-t_1-z_2-t_2+2z_3+2t_3})}{(1-q^{z_1-2z_2+z_3})(1-q^{-z_1-z_2+2z_3})}\\ & \cdot\frac{(1-q^{z_1+t_1-z_2-t_2})(1-q^{-z_1-t_1+z_3+t_3})(1-q^{-z_2-t_2+z_3+t_3})}{(1-q^{z_1-z_2})(1-q^{-z_1+z_3})(1-q^{-z_2+z_3})}\\ & \cdot q^{\sum_{1\le j<k\le 3}\{4(z_j-z_k)(t_j-t_k)+2(t_j-t_k)^2\}}\Bigg\} \end{aligned} \tag{1.6}$$

$$=[q]_\infty^2\prod_{i=1}^{3}[q^{1+3z_i-z_1-z_2-z_3}]_\infty[q^{1-3z_i+z_1+z_2+z_3}]_\infty\prod_{\substack{1\le j,k\le 3\\ j\ne k}}[q^{1+z_j-z_k}]_\infty.$$

Proof. Let $a_i \to \infty$ and $b_i \to 0$ for $i = 1, 2$ in (1.3). Then the argument in Theorem 7.1 of [10] shows that (1.6) is equivalent to the Macdonald identity (Theorem 8.7 of [15]) for G_2.

In § 3 we will prove the limiting case ($q \to 1$) of Theorem 1.1.

THEOREM 1.7. *Let $z = (z_1, z_2, z_3) \in \mathbb{C}^3$ and $\alpha_i, \beta_i \in \mathbb{C}$ for $i = 1, 2$. Let L be defined as in Theorem* 1.1. *Let*

$$\gamma_i = \begin{cases} \alpha_i & \text{for } i = 1, 2, \\ 1 - \beta_{i-2} & \text{for } i = 3, 4. \end{cases} \tag{1.8}$$

Then we have

$$\begin{aligned} &\sum_{t=(t_1,t_2,t_3)\in L} \left\{ \left(\frac{-2z_1 - 2t_1 + z_2 + t_2 + z_3 + t_3}{-2z_1 + z_2 + z_3} \right) \right. \\ &\quad \cdot \left(\frac{z_1 + t_1 - 2z_2 - 2t_2 + z_3 + t_3}{z_1 - 2z_2 + z_3} \right) \left(\frac{-z_1 - t_1 - z_2 - t_2 + 2z_3 + 2t_3}{-z_1 - z_2 + 2z_3} \right) \\ &\quad \cdot \left(\frac{z_1 + t_1 - z_2 - t_2}{z_1 - z_2} \right) \left(\frac{-z_1 - t_1 + z_3 + t_3}{-z_1 + z_3} \right) \left(\frac{-z_2 - t_2 + z_3 + t_3}{-z_2 + z_3} \right) \\ &\quad \left. \cdot \prod_{i=1}^{2} \prod_{\substack{1 \le j,k \le 3 \\ j \ne k}} \frac{(\alpha_i + z_j - z_k)_{t_j - t_k}}{(\beta_i + z_j - z_k)_{t_j - t_k}} \right\} \\ &= \frac{\Gamma\left(1 - 2\sum_{i=1}^{4} \gamma_i\right) \prod_{i=1}^{4} \Gamma(1 - \gamma_i) \prod_{i=1}^{4} \prod_{\substack{1 \le j,k \le 3 \\ j \ne k}} \Gamma(1 - \gamma_i + z_j - z_k)}{\Gamma\left(1 - \sum_{i=1}^{4} \gamma_i\right) \prod_{1 \le i \le j \le 4} \Gamma(1 - \gamma_i - \gamma_j) \prod_{1 \le i < j < k \le 4} \Gamma(1 - \gamma_i - \gamma_j - \gamma_k)} \\ &\quad \cdot \left\{ \prod_{i=1}^{3} \Gamma(1 + 3z_i - z_1 - z_2 - z_3) \right. \\ &\quad \left. \cdot \Gamma(1 - 3z_i + z_1 + z_2 + z_3) \prod_{\substack{1 \le j,k \le 3 \\ j \ne k}} \Gamma(1 + z_j - z_k) \right\}^{-1}, \end{aligned} \tag{1.9}$$

where $\operatorname{Re}(2\sum_{i=1}^{2}(\beta_i - \alpha_i)) > 3$ *for convergence and the poles of both sides of* (1.9) *are avoided.*

We hope that Theorem 1.1 will have application in the theory of partitions of multidimensional analogues (see [1], [2]). It seems likely that families of orthogonal polynomials in two variables will be associated to these summation theorems (see [3], [4], [8], and [19]).

The results of this paper provide evidence that there may be summation theorems for very-well-poised hypergeometric series associated to the exceptional simple Lie algebras over $\mathbb{C}$ or, alternatively, to affine root systems of exceptional type. Limiting cases of these summation theorems should give the Macdonald identities for affine root systems of exceptional type.

There is a problem in extending the methods of this paper to the other exceptional Lie algebras (or affine root systems). In order to find the corresponding difference equations there need to be sufficiently large numbers of numerator and denominator parameters a_i and b_i. For these other exceptional Lie algebras (or affine root systems) there does not appear to be a sufficient number of parameters to produce the required difference equation.

2. Proof of Theorem 1.1. We begin the proof of identity (1.3) by first showing that the series on the left-hand side of (1.3) converges whenever $|q^{-3}\prod_{i=1}^{2}(b_i/a_i)^2|<1$ and none of the denominators vanish. With notation as in Theorem 1.1, let Φ be the root system of type G_2 (see [7, App.]) with a partial ordering and W the Weyl group of G_2, $W\approx\mathfrak{D}_6$, the dihedral group of order 12. Let $\rho=\frac{1}{2}\sum_{\alpha>0}\alpha$. Using the Weyl denominator formula for G_2 (see [12]), the left-hand side of (1.3) can be rewritten as

$$\sum_{t\in L}\left\{q^{-\rho(t)}\prod_{\alpha>0}\frac{(1-q^{\alpha(z+t)})}{(1-q^{\alpha(z)})}\prod_{i=1}^{2}\prod_{\substack{1\leqq j,k\leqq 3\\ j\neq k}}\frac{[a_iq^{z_j-z_k}]_{t_j-t_k}}{[b_iq^{z_j-z_k}]_{t_j-t_k}}\right\}$$

$$=\left(\prod_{\alpha>0}(1-q^{-\alpha(z)})\right)^{-1}\sum_{t\in L}\sum_{w\in W}\tag{2.1}$$

$$\cdot\left\{\varepsilon(w)q^{-w\rho(t+z)}\prod_{i=1}^{2}\prod_{\substack{1\leqq j,k\leqq 3\\ j\neq k}}\frac{[a_iq^{z_j-z_k}]_{t_j-t_k}}{[b_iq^{z_j-z_k}]_{t_j-t_k}}\right\}$$

where $\varepsilon(w)$ is the sign of the Weyl group element $w\in W$ and $t=(t_1,t_2,t_3)$ and $z=(z_1,z_2,z_3)$ in the notation of Theorem 1.1.

Let $t\in L$ and $t=n_1\varphi_1+n_2\varphi_2+n_3\varphi_3$ for some choice of n_1, n_2, $n_3\in\mathbb{Z}$. Suppose

$$n_{\sigma(1)}\geqq n_{\sigma(2)}\geqq n_{\sigma(3)},$$

where σ is a permutation of $\{1,2,3\}$. Let $l_1=n_{\sigma(1)}-n_{\sigma(2)}$ and $l_2=n_{\sigma(2)}-n_{\sigma(3)}$. Since $\varphi_1+\varphi_2+\varphi_3=0$, then we have

$$t=l_1\varphi_{\sigma(1)}+l_2(-\varphi_{\sigma(3)})\tag{2.2}$$

with l_1, $l_2\geqq 0$. It follows that

$$\prod_{i=1}^{2}\prod_{\substack{1\leqq j,k\leqq 3\\ j\neq k}}\frac{[a_iq^{z_j-z_k}]_{t_j-t_k}}{[b_iq^{z_j-z_k}]_{t_j-t_k}}$$

$$=\prod_{i=1}^{2}\left\{\frac{[a_iq^{z_{\sigma(1)}-z_{\sigma(2)}}]_{l_1}[a_iq^{z_{\sigma(2)}-z_{\sigma(1)}}]_{-l_1}}{[b_iq^{z_{\sigma(1)}-z_{\sigma(2)}}]_{l_1}[b_iq^{z_{\sigma(2)}-z_{\sigma(1)}}]_{-l_1}}\right.\tag{2.3}$$

$$\left.\cdot\frac{[a_iq^{z_{\sigma(2)}-z_{\sigma(3)}}]_{l_2}[a_iq^{z_{\sigma(3)}-z_{\sigma(2)}}]_{-l_2}[a_iq^{z_{\sigma(1)}-z_{\sigma(3)}}]_{l_1+l_2}[a_iq^{z_{\sigma(3)}-z_{\sigma(1)}}]_{-l_1-l_2}}{[b_iq^{z_{\sigma(2)}-z_{\sigma(3)}}]_{l_2}[b_iq^{z_{\sigma(3)}-z_{\sigma(2)}}]_{-l_2}[b_iq^{z_{\sigma(1)}-z_{\sigma(3)}}]_{l_1+l_2}[b_iq^{z_{\sigma(3)}-z_{\sigma(1)}}]_{-l_1-l_2}}\right\}.$$

From (2.1)-(2.3) we see that the left-hand side of (1.3) converges absolutely if we can show for all $\sigma\in S_3$ and $w\in W$ the absolute convergence of the following series:

$$\sum_{\substack{t=l_1\varphi_{\sigma(1)}-l_2\varphi_{\sigma(3)}\\ l_1,l_2\geqq 0}}q^{-w\rho(t+z)}\prod_{i=1}^{2}\left\{\frac{[a_iq^{z_{\sigma(1)}-z_{\sigma(2)}}]_{l_1}[a_iq^{z_{\sigma(2)}-z_{\sigma(1)}}]_{-l_1}}{[b_iq^{z_{\sigma(1)}-z_{\sigma(2)}}]_{l_1}[b_iq^{z_{\sigma(2)}-z_{\sigma(1)}}]_{-l_1}}\frac{[a_iq^{z_{\sigma(2)}-z_{\sigma(3)}}]_{l_2}}{[b_iq^{z_{\sigma(2)}-z_{\sigma(3)}}]_{l_2}}\right.\tag{2.4}$$

$$\left.\cdot\frac{[a_iq^{z_{\sigma(3)}-z_{\sigma(2)}}]_{-l_2}}{[b_iq^{z_{\sigma(3)}-z_{\sigma(2)}}]_{-l_2}}\frac{[a_iq^{z_{\sigma(1)}-z_{\sigma(3)}}]_{l_1+l_2}}{[b_iq^{z_{\sigma(1)}-z_{\sigma(3)}}]_{l_1+l_2}}\frac{[a_iq^{z_{\sigma(3)}-z_{\sigma(1)}}]_{-l_1-l_2}}{[b_iq^{z_{\sigma(3)}-z_{\sigma(1)}}]_{-l_1-l_2}}\right\}.$$

We will use the ratio test to show the convergence of (2.4).

Let d_u be the $t=u$ term in the series (2.4). Consider

$$\frac{d_{(t+\varphi_{\sigma(1)})}}{d_t}=q^{-w\rho(\varphi_{\sigma(1)})}\prod_{i=1}^{2}\left\{\left(\frac{1-a_iq^{z_{\sigma(1)}-z_{\sigma(2)}+l_1}}{1-b_iq^{z_{\sigma(1)}-z_{\sigma(2)}+l_1}}\right)\left(\frac{1-b_iq^{z_{\sigma(2)}-z_{\sigma(1)}-l_1-1}}{1-a_iq^{z_{\sigma(2)}-z_{\sigma(1)}-l_1-1}}\right)\right.$$

$$\left.\cdot\left(\frac{1-a_iq^{z_{\sigma(1)}-z_{\sigma(3)}+l_1+l_2}}{1-b_iq^{z_{\sigma(1)}-z_{\sigma(3)}+l_1+l_2}}\right)\left(\frac{1-b_iq^{z_{\sigma(3)}-z_{\sigma(1)}-l_1-l_2-1}}{1-a_iq^{z_{\sigma(3)}-z_{\sigma(1)}-l_1-l_2-1}}\right)\right\}\tag{2.5}$$

$$= q^{-w\rho(\varphi_{\sigma(1)})}\left(\frac{b_1b_2}{a_1a_2}\right)^2 M(l_1, l_2),$$

where

$$M(l_1, l_2) = \prod_{i=1}^{2}\left\{\left(\frac{1-a_iq^{z_{\sigma(1)}-z_{\sigma(2)}+l_1}}{1-b_iq^{z_{\sigma(1)}-z_{\sigma(2)}+l_1}}\right)\left(\frac{1-b_i^{-1}q^{z_{\sigma(1)}-z_{\sigma(2)}+l_1+1}}{1-a_i^{-1}q^{z_{\sigma(1)}-z_{\sigma(2)}+l_1+1}}\right)\right.$$
$$\left.\cdot\left(\frac{1-a_iq^{z_{\sigma(1)}-z_{\sigma(3)}+l_1+l_2}}{1-b_iq^{z_{\sigma(1)}-z_{\sigma(3)}+l_1+l_2}}\right)\left(\frac{1-b_i^{-1}q^{z_{\sigma(1)}-z_{\sigma(3)}+l_1+l_2+1}}{1-a_i^{-1}q^{z_{\sigma(1)}-z_{\sigma(3)}+l_1+l_2+1}}\right)\right\}.$$

There is a similar formula for $d_{t-\varphi_{\sigma(3)}}/d_t$.

By choosing l_1 and l_2 sufficiently large, we can make $|M(l_1, l_2)|$ as close to 1 as desired, with a corresponding statement for $d_{t-\varphi_{\sigma(3)}}/d_t$. Hence by the ratio test for multiple series, the series (2.4) will converge if

$$|q^{-w\rho(\varphi_{\sigma(1)})}(b_1b_2)^2/(a_1a_2)^2| < 1, \qquad |q^{w\rho(\varphi_{\sigma(3)})}(b_1b_2)^2/(a_1a_2)^2| < 1.$$

Now $\rho = -\varepsilon_1 - 2\varepsilon_2 + 3\varepsilon_3$, where ε_i is the standard basis vector in $\mathbb{C}^3$ and $w\rho = \pm(-\varepsilon_{\tau(1)} - 2\varepsilon_{\tau(2)} + 3\varepsilon_{\tau(3)})$ for some permutation $\tau \in S_3$. It follows that for all $w \in W$ and $\sigma \in S_3$ we have

$$|q^{-w\rho(\varphi_{\sigma(1)})}| \leqq q^{-3}, \qquad |q^{w\rho(\varphi_{\sigma(3)})}| \leqq q^{-3}.$$

Hence the original series on the left-hand side of identity (1.3) will converge whenever

$$|q^{-3}(b_1b_2)^2/(a_1a_2)^2| < 1$$

and denominators do not vanish.

To prove Theorem 1.1 we will show that both sides of (1.3) satisfy a q-difference equation in the parameters c_i, $1 \leqq i \leqq 4$. If for some $i \neq j$, c_ic_j equals 1, we will show that (1.3) reduces to an identity equivalent to a special case of the summation theorem for multilateral hypergeometric series that are very well poised on the Lie algebra of type A_1 [9], [10]. By means of this special evaluation of (1.3) and the q-difference equation, the general result follows by induction and analytic continuation of the parameters c_i.

DEFINITION 2.6. If $f(c_j)$ is a function of the variable c_j, then

$$(R_jf)(c_j) = f(c_jq).$$

LEMMA 2.7. *Let $f(c_1, \cdots, c_4, z) = f$ be the left-hand side of the right-hand side of* (1.3). *Then f satisfies the q-difference equation* (*where we assume that* $\prod_{1\leqq i<j\leqq 4}(c_i - c_j) \times (c_ic_j - 1) \neq 0$):

$$\sum_{i=1}^{4}\prod_{\substack{l=1\\ l\neq i}}^{4}\frac{c_i}{(c_l - c_i)(c_lc_i - 1)}\prod_{\substack{1\leqq j,k\leqq 3\\ j\neq k}}(1 - c_lq^{z_j - z_k})(R_lf) = f. \tag{2.8}$$

Proof. That the left-hand side of (1.3) satisfies the q-difference equation (2.8) is essentially the same proof as in Lemmas 5.3. and 6.3 of [10]. We give the proof below.

For any i, $1 \leqq i \leqq 4$, we define

$$d_i(m) = c_i^m + c_i^{6-m} \quad \text{for } 0 \leqq m < 3, \tag{2.9a}$$

$$d_i(3) = c_i^3. \tag{2.9b}$$

Consider the determinant

$$\sum_{\sigma\in S_4}\varepsilon(\sigma)\prod_{i=0}^{3}d_{\sigma(i+1)}(i) = \left(\prod_{i=1}^{4}c_i^3\right)\sum_{(\sigma,t)\in S_4\times T}\left\{\varepsilon(\sigma)\prod_{i=1}^{4}c_{\sigma(i)}^{(l+1-i)t_i}\right\}, \tag{2.10}$$

where $T=\{(t_1,\cdots,t_4)|t_i\in\{1,-1\}$ for $1\leqq i\leqq 4$ and $\sum_{i=1}^4 t_i\equiv 0 \pmod 4\}$. From the Weyl denominator formula for the root system of type D_4 [12], we find that the right hand side of (2.10) equals

$$\begin{aligned}(2.11)\qquad &\prod_{i=1}^4 c_i^3 \prod_{1\leqq i<j\leqq 4}[((c_i/c_j)^{1/2}-(c_j/c_i)^{1/2})((c_ic_j)^{1/2}-(c_ic_j)^{-1/2})]\\ &=\prod_{1\leqq i<j\leqq 4}(c_i-c_j)(c_ic_j-1).\end{aligned}$$

More generally, we have that the determinant

$$(2.12)\qquad \sum_{\sigma\in S_4}\left\{\varepsilon(\sigma)d_{\sigma(1)}(m)\prod_{i=1}^3 d_{\sigma(i+1)}(i)\right\}$$

vanishes unless $m=0$, in which case it equals $\prod_{1\leqq i<j\leqq 4}(c_i-c_j)(c_ic_j-1)$. By expanding the determinant (2.12) along the first column and using the Weyl denominator formula for the root system of type D_3, we find (2.12) equals

$$(2.13)\qquad \sum_{l=1}^4(-1)^{l-1}d_l(m)\prod_{\substack{i=1\\ i\neq l}}^4(c_i)\prod_{\substack{1\leqq i<j\leqq 4\\ i,j\neq l}}(c_i-c_j)(c_ic_j-1).$$

Dividing (2.13) by $\prod_{1\leqq i<j\leqq 4}(c_i-c_j)(c_ic_j-1)$, we obtain the identity

$$(2.14)\qquad \sum_{l=1}^4 d_l(m)\prod_{\substack{i=1\\ i\neq l}}^4\frac{c_i}{(c_l-c_i)(c_lc_i-1)}=\begin{cases}0 & \text{if } 1\leqq m\leqq 3,\\ 1 & \text{if } m=0.\end{cases}$$

Now let ψ denote the left-hand side of (1.3). For any l, $1\leqq l\leqq 4$, consider

$$\begin{aligned}(2.15)\qquad R_l(\psi)=\sum_{t\in L}&\prod_{\substack{1\leqq j,k\leqq 3\\ j\neq k}}\left(\frac{1-c_lq^{z_j+t_j-z_k-t_k}}{1-c_lq^{z_j-z_k}}\right)\\ &\cdot q^{-\rho(t)}\prod_{\alpha>0}\left(\frac{1-q^{\alpha(z+t)}}{1-q^{\alpha(z)}}\right)\prod_{i=1}^2\prod_{\substack{1\leqq j,k\leqq 3\\ j\neq k}}\frac{[a_iq^{z_j-z_k}]_{t_j-t_k}}{[b_iq^{z_j-z_k}]_{t_j-t_k}}\Bigg\},\end{aligned}$$

with notation as in (2.1):

$$\begin{aligned}R_l(\psi)=\sum_{t\in L}\Bigg\{&\sum_{m=0}^3(-1)^m[d_l(m)e_m(q^{z_j+t_j-z_k-t_k})]/\prod_{\substack{1\leqq j,k\leqq 3\\ j\neq k}}(1-c_lq^{z_j-z_k})\\ &\cdot q^{-\rho(t)}\prod_{\alpha>0}\left(\frac{1-q^{\alpha(z+t)}}{1-q^{\alpha(z)}}\right)\prod_{i=1}^2\prod_{\substack{1\leqq j,k\leqq 3\\ j\neq k}}\frac{[a_iq^{z_j-z_k}]_{t_j-t_k}}{[b_iq^{z_j-z_k}]_{t_j-t_k}}\Bigg\},\end{aligned}$$

where $e_m(q^{z_j+t_j-z_k-t_k})$ is the elementary symmetric function of degree m in the variables $q^{z_j+t_j-z_k-t_k}$, $1\leqq j\neq k\leqq 3$. Note that $e_m(q^{z_j+t_j-z_k-t_k})=e_{6-m}(q^{z_j+t_j-z_k-t_k})$ for $0\leqq m\leqq 3$ and $e_0(q^{z_j+t_j-z_k-t_k})\equiv 1$.

It follows that

$$\begin{aligned}&\sum_{l=1}^4\left\{\prod_{\substack{i=1\\ i\neq l}}^4\frac{c_i}{(c_l-c_i)(c_lc_i-1)}\prod_{\substack{1\leqq j,k\leqq 3\\ j\neq k}}(1-c_lq^{z_j-z_k})(R_l\psi)\right\}\\ &\quad=\sum_{t\in L}\sum_{m=0}^3\sum_{l=1}^4\left\{d_l(m)\prod_{\substack{i=1\\ i\neq l}}^4\frac{c_i}{(c_l-c_i)(c_lc_i-1)}(-1)^me_m(q^{z_j+t_j-z_k-t_k})\right.\end{aligned}$$

$$\cdot q^{-\rho(t)} \prod_{\alpha>0} \left(\frac{1-q^{\alpha(z+t)}}{1-q^{\alpha(z)}} \right) \prod_{i=1}^{2} \prod_{\substack{1\leqq j,k\leqq 3 \\ j\neq k}} \frac{[a_i q^{z_j-z_k}]_{t_j-t_k}}{[b_i q^{z_j-z_k}]_{t_j-t_k}} \Bigg\} \tag{2.16}$$

$$= \sum_{t\in L} q^{-\rho(t)} \prod_{\alpha>0} \left(\frac{1-q^{\alpha(z+t)}}{1-q^{\alpha(z)}} \right) \prod_{i=1}^{2} \prod_{\substack{1\leqq j,k\leqq 3 \\ j\neq k}} \frac{[a_i q^{z_j-z_k}]_{t_j-t_k}}{[b_i q^{z_j-z_k}]_{t_j-t_k}}$$

$$= \psi$$

by (2.14). Thus the left-hand side of (1.3) satisfies the q-difference equation (2.8).

Let Q denote the right-hand side of (1.3). For any l, $1\leqq l\leqq 4$, consider

$$R_l(Q) = Q\cdot(1-q^{-1}c_l^{-2}) \prod_{i=1}^{4} (1-c_l^{-1}c_i^{-1})$$

$$\cdot \frac{\displaystyle\prod_{j,k\neq l,1\leqq j<k\leqq 4} (1-c_l^{-1}c_j^{-1}c_k^{-1})(1-\prod_{i=1}^{4} c_i^{-1})}{(1-c_l^{-1})(1-q^{-1}\prod_{i=1}^{4} c_i^{-2})(1-\prod_{i=1}^{4} c_i^{-2}) \displaystyle\prod_{1\leqq j,k\leqq 3, j\neq k} (1-c_l^{-1}q^{z_j-z_k})}. \tag{2.17}$$

Thus, to show that Q satisfies the q-difference equation (2.8), we are reduced to verifying the identity

$$\sum_{l=1}^{4} \frac{c_l^3}{\displaystyle\prod_{i=1,i\neq l}^{4} (c_l-c_i)} (1-q^{-1}c_l^{-2})(1+c_l^{-1}) \prod_{\substack{1\leqq j<k\leqq 4 \\ j,k\neq l}} (1-c_l^{-1}c_j^{-1}c_k^{-1})$$

$$= \left(1-q^{-1}\prod_{i=1}^{4} c_i^{-2}\right)\left(1+\prod_{i=1}^{4} c_i^{-1}\right). \tag{2.18}$$

If $\{X_1, \cdots, X_n\}$ is a set of variables, then

$$e_k(X_1, \cdots, X_n) = \sum_{1\leqq i_1<i_2<\cdots<i_k\leqq n} X_{i_1}X_{i_2}\cdots X_{i_k}$$

is the elementary symmetric function of degree k. We have

$$\prod_{\substack{1\leqq j<k<4 \\ j,k\neq l}} (1-c_l^{-1}c_j^{-1}c_k^{-1}) = \sum_{m=0}^{3} (-1)^m c_l^{-m} e_m(c_j^{-1}c_k^{-1}), \tag{2.19}$$

where $e_m(c_j^{-1}c_k^{-1})$ is the elementary symmetric function of degree m in the variables $\{c_j^{-1}c_k^{-1} \mid 1\leqq j<k\leqq 4; j, k\neq l\}$. We can also write

$$\begin{aligned} &e_0(c_j^{-1}c_k^{-1}) = 1, \\ &e_1(c_j^{-1}c_k^{-1}) = e_2(c_j^{-1}), \\ &e_2(c_j^{-1}c_k^{-1}) = e_1(c_j^{-1}) \prod_{\substack{i=1 \\ i\neq l}}^{4} (c_i)^{-1}, \\ &e_3(c_j^{-1}c_k^{-1}) = \prod_{\substack{i=1 \\ i\neq l}}^{4} (c_i)^{-2}, \end{aligned} \tag{2.20}$$

where the elementary symmetric functions on the right-hand sides of (2.20) are in the variables $\{c_j^{-1} \mid 1\leqq j\leqq 4, j\neq l\}$. Note that $e_3(c_j^{-1}) = \prod_{i=1;i\neq l}^{4} (c_i)^{-1}$ and $e_4(c_j^{-1}) = 0$. We also have

$$c_l^{-1}e_m(c_j^{-1}) = e_{m+1}(c_1^{-1}, c_2^{-1}, c_3^{-1}, c_4^{-1}) - e_{m+1}(c_j^{-1}), \tag{2.21}$$

where $0 \leqq m \leqq 3$, and $e_m(c_j^{-1})$ and $e_{m+1}(c_j^{-1})$ are functions in the variables $\{c_j^{-1} \mid 1 \leqq j \leqq 4, j \neq l\}$. It follows that

$$\sum_{m=0}^{3} (-1)^m c_l^{-m} e_m(c_j^{-1}c_k^{-1}) = 1 - e_3 + c_l e_4 + e_2 e_4 - c_l e_3 e_4 + c_l^2 e_4^2 - c_l^{-1} e_4^2, \tag{2.22}$$

where the symmetric functions on the right-hand side of (2.22) are in the variables $c_1^{-1}, c_2^{-1}, c_3^{-1}, c_4^{-1}$ and on the left-hand side of (2.22) are in the variables $\{c_j^{-1}c_k^{-1} \mid 1 \leqq j < k \leqq 4; j, k \neq l\}$.

To calculate the left-hand side of (2.18), we will need an important lemma due to Louck and Biedenharn [14] (see also [11]).

LEMMA 2.23. *Let $X_1, \cdots, X_n$ be arbitrary distinct complex numbers, of any non-negative integer, and $h_l(X)$ the lth homogeneous symmetric function of $X_1, \cdots, X_n$. We then have*

$$\sum_{i=1}^{n} (X_i)^q \prod_{\substack{s=1 \\ s \neq i}}^{n} (X_s - X_i)^{-1} = (-1)^{n-1} h_{q-n+1}(X).$$

Note that $h_k(X) \equiv 0$ if $k < 0$ and $h_0(X) \equiv 1$.

Using the identity

$$c_l^3 \prod_{\substack{i=1 \\ i \neq l}}^{4} (c_l - c_i)^{-1} = -\prod_{\substack{i=1 \\ i \neq l}}^{4} \left(\frac{c_i^{-1}}{c_l^{-1} - c_i^{-1}} \right), \tag{2.24}$$

we substitute identities (2.19) and (2.22) into the left-hand side of (2.18). Thus the left-hand side of (2.18) equals

$$\begin{aligned}
&\sum_{l=1}^{4} \frac{c_l^3}{\prod_{i=1; i=l}^{4} (c_l - c_i)} (1 - q^{-1}c_l^{-2})(1 + c_l^{-1})\{1 - e_3 + c_l e_4 + e_2 e_4 - c_l e_3 e_4 + c_l^2 e_4^2\} \\
&\quad - \sum_{l=1}^{4} \frac{c_l^3}{\prod_{i=1, i \neq l}^{4} (c_l - c_i)} (1 + c_l^{-1} - q^{-1}c_l^{-2}) c_l^{-1} e_4^2 + \sum_{l=1}^{4} \frac{(-q^{-1}c_l^{-3}) e_4^3}{\prod_{i=1, i \neq l}^{4} (c_l^{-1} - c_i^{-1})} \\
&= 1 - e_3(c_i^{-1}) + h_1(c_i) e_4(c_i^{-1}) + e_4(c_i^{-1}) + e_2(c_i^{-1}) e_4(c_i^{-1}) \\
&\quad - h_1(c_i) e_3(c_i^{-1}) e_4(c_i^{-1}) - e_3(c_i^{-1}) e_4(c_i^{-1}) + h_2(c_i) e_4^2(c_i^{-1}) \\
&\quad + h_1(c_i) e_4^2(c_i^{-1}) - q^{-1} e_4^2(c_i^{-1}) - q^{-1} e_4^3(c_i^{-1}),
\end{aligned} \tag{2.25}$$

where the symmetric functions e_k on the left-hand side of (2.25) are in the variables c_i^{-1}, $1 \leqq i \leqq 4$, and on the right-hand side of (2.25) the symmetric functions $h_k(c_i)$ and $e_k(c_i^{-1})$ are in the variables c_i and c_i^{-1}, $1 \leqq i \leqq 4$, respectively.

Note that $h_1(c_i) e_4(c_i^{-1}) = e_3(c_i^{-1})$, with notation as above. A simple case of the Jacobi-Trudi identity [16] is $h_2(c_i) = e_1^2(c_i) - e_2(c_i)$. Since $h_1(c_i) = e_1(c_i)$ and $e_2(c_i) e_4(c_i^{-1}) = e_2(c_i^{-1})$, this implies $h_2(c_i) e_4^2(c_i^{-1}) = e_3^2(c_i^{-1}) - e_2(c_i^{-1}) e_4(c_i^{-1})$, with notation as above. Hence, the right-hand side of (2.25) equals

$$\begin{aligned}
&1 - e_3 + e_3 + e_4 + e_2 e_4 - e_3^2 - e_3 e_4 + e_3^2 - e_2 e_4 + e_3 e_4 - q^{-1} e_4^2 - q^{-1} e_4^3 \\
&\quad = 1 + e_4 - q^{-1} e_4^2 - q^{-1} e_4^3,
\end{aligned} \tag{2.26}$$

where all symmetric functions are in the variables c_i^{-1}, $1 \leqq i \leqq 4$. The right-hand side of (2.26) equals the right-hand side of (2.18). This completes the proof of (2.18), and also the proof of Lemma 2.7. In other words, both sides of (1.3) satisfy the same q-difference equation (2.8).

To complete the proof of (1.3) we need to consider a special evaluation of the series on the left-hand side of (1.3).

LEMMA 2.27. *Assume* $|q^{-3}\prod_{i=1}^{2}(b_i/a_i)^2| < 1$ *and none of the denominators on both sides of* (1.3) *vanish; then* (1.3) *is valid whenever* $c_i c_j = 1$ *for some* $i \neq j$, $1 \leqq i, j \leqq 4$.

Proof. Let $p_j : \mathbb{C}^3 \to \mathbb{C}$, $1 \leqq j \leqq 3$, be linear functionals such that for all $X = (X_1, X_2, X_3) \in \mathbb{C}^3$ we have $p_1(X) = X_1 - X_2$, $p_2(X) = X_2 - X_3$, and $p_3(X) = X_3 - X_1$. For $z = (z_1, z_2, z_3)$ as in (1.3), define $p_1(z) = z_1 - z_2 = u_1$, $p_2(z) = z_2 - z_3 = u_2$, and $p_3(z) = z_3 - z_1 = u_3$. Similarly, for $t = (t_1, t_2, t_3) \in L$, define $p_1(t) = t_1 - t_2 = y_1$, $p_2(t) = t_2 - t_3 = y_2$, and $p_3(t) = t_3 - t_1 = y_3$. For the generators φ_1, φ_2, φ_3 of L, we have the following:

$$p_1(\varphi_1) = 1, \quad p_1(\varphi_2) = -1, \quad p_1(\varphi_3) = 0,$$

$$p_2(\varphi_1) = 0, \quad p_2(\varphi_2) = 1, \quad p_2(\varphi_3) = -1,$$

$$p_3(\varphi_1) = -1, \quad p_3(\varphi_2) = 0, \quad p_3(\varphi_3) = 1.$$

Hence for $t \in L$ and $t = n_1\varphi_1 + n_2\varphi_2 + n_3\varphi_3$, we have

$$p_1(t) = n_1 - n_2 = y_1, \quad p_2(t) = n_2 - n_3 = y_2, \quad p_3(t) = n_3 - n_1 = y_3.$$

If $t = n_1'\varphi_1 + n_2'\varphi_2 + n_3'\varphi_3$, then $n_1 - n_2 = n_1' - n_2' = y_1$, $n_2 - n_3 = n_2' - n_3' = y_2$, and $n_3 - n_1 = n_3' - n_1' = y_3$. Conversely, given any triple $(y_1, y_2, y_3) \in \mathbb{Z}^3$ such that $y_1 + y_2 + y_3 = 0$, let $t = -y_3\varphi_1 + y_2\varphi_2$. Then we have $p_1(t) = -y_3 - y_2 = y_1$, $p_2(t) = y_2$, and $p_3(t) = y_3$. It follows that there is a one-to-one correspondence between elements $t \in L$ and triples $(y_1, y_2, y_3) \in \mathbb{Z}^3$ such that $y_1 + y_2 + y_3 = 0$.

Consequently, in the above notation we may rewrite the left-hand side of (1.3) as follows:

$$\begin{aligned}\sum_{\substack{y_1,y_2,y_3=-\infty \\ y_1+y_2+y_3=0}}^{\infty} \Bigg\{ & q^{y_1+3y_2}\left(\frac{1-q^{u_3+y_3-u_1-y_1}}{1-q^{u_3-u_1}}\right) \\ & \cdot\left(\frac{1-q^{u_1+y_1-u_2-y_2}}{1-q^{u_1-u_2}}\right)\left(\frac{1-q^{u_3+y_3-u_2-y_2}}{1-q^{u_3-u_2}}\right)\left(\frac{1-q^{u_1+y_1}}{1-q^{u_1}}\right) \\ & \cdot\left(\frac{1-q^{u_3+y_3}}{1-q^{u_3}}\right)\left(\frac{1-q^{-u_2-y_2}}{1-q^{-u_2}}\right)\prod_{i=1}^{4}\prod_{j=1}^{3}\frac{[c_i q^{u_j}]_{y_j}}{[qc_i^{-1}q^{u_j}]_{y_j}}\Bigg\}.\end{aligned} \tag{2.28}$$

Now let us assume for some $i \neq j$, $1 \leqq i, j \leqq 4$, that $c_i = c_j^{-1} = q^{u_i}$. Suppose that $c_4 = c_3^{-1} = q^{u_3}$. The proof is similar in the other cases. Then (2.28) reduces to the following:

$$\sum_{\substack{y_1,y_2=-\infty \\ y_1+y_2=0}}^{\infty} \left\{\left(\frac{q^{u_1+y_1}-q^{u_2+y_2}}{q^{u_1}-q^{u_2}}\right)\prod_{i,j=1}^{2}\frac{[c_i q^{u_j}]_{y_j}}{[qc_i^{-1}q^{u_j}]_{y_j}}\right\}. \tag{2.29}$$

Expression (2.29) is a basic hypergeometric series that is very well poised on A_1 (or $SU(2)$), i.e., a classical ${}_6\psi_6$ series. Its sum is evaluated as a special case of theorem 1.15 of [9]. (This case is simply a reformulation of Bailey's ${}_6\psi_6$ sum [5].) We find that

(2.29) equals

$$\frac{[q]_\infty[q^{1+u_1-u_2}]_\infty[q^{1+u_2-u_1}]_\infty\prod_{i,k=1}^{2}[qc_i^{-1}c_k^{-1}]_\infty}{[q(c_1c_2)^{-2}]_\infty\prod_{i,k=1}^{2}[qc_i^{-1}q^{u_k}]_\infty[qc_i^{-1}q^{-u_k}]_\infty} \tag{2.30}$$

$$\cdot\,[q^{1+u_1+u_2}(c_1c_2)^{-1}]_\infty[q^{1-u_1-u_2}(c_1c_2)^{-1}]_\infty.$$

Expression (2.30) equals the right-hand side of (1.3) in the case $c_4 = c_3^{-1} = q^{u_3}$.

We will now eliminate the restriction that $c_4 = c_3^{-1} = q^{u_3}$ and simply require that $c_4 = c_3^{-1}$. Let ψ and Q be the left- and right-hand sides, respectively, of (1.3). Define $\Phi(z) = Q^{-1}\psi$ as a function of $z = (z_1, z_2, z_3)$, where $z_1 + z_2 + z_3 = 0$. We have the following lemma.

LEMMA 2.31 [10, Lemma 3.15]. *$\Phi(z)$ is a constant for all z such that $z_1 + z_2 + z_3 = 0$.*

Proof. This lemma is an immediate consequence of Lemma 3.15 of [10]. It is only necessary to verify that

$$\prod_{\alpha>0}\alpha(z)^2 = 2\sum_{\substack{1\le j,k\le 3\\ j\ne k}}(z_j - z_k)^2$$

with notation as in (2.1). This can be done without much difficulty.

As a consequence of Lemma 2.31, it follows that since (1.3) is valid for $c_4 = c_3^{-1} = q^{u_3}$, it must also be valid for $c_4 = c_3^{-1}$ regardless of u_3. More generally, (1.3) is valid whenever $c_ic_j = 1$ for some $i \ne j$, $1 \le i, j < 4$. This completes the proof of Lemma 2.27.

Using the special evaluation in Lemma 2.27 as a starting point, we will prove (1.3) by induction. Assume that $c_1c_3 = q^{-k_1}$ and $c_2c_4 = q^{-k_2}$, where k_1 and k_2 are nonnegative integers. Furthermore, assume temporarily that $c_ic_j^{-1} \ne q^l$ for any integer l and all i, j such that $1 \le i \ne j \le 4$. Let $k = k_1 + k_2$. The assumption $k \ge 0$ guarantees the convergence of the series on the left-hand side of (1.3). Now assume that (1.3) is valid for all such c_i with $0 \le k \le N$ for some integer N. Consider the case $k = N+1$. If $k_1 = 0$ or $k_2 = 0$, then (1.3) is a consequence of Lemma 2.27. If both k_1 and k_2 are positive, then (1.3) follows from the induction hypothesis and Lemma 2.7. By induction this proves that (1.3) is valid whenever $c_ic_{i+2} = q^{-k_i}$, $i = 1,2$, for all nonnegative integers k_i (assuming no denominators vanish).

Both sides of (1.3) are analytic functions in the variables c_i^{-1}, $1 \le i \le 4$. In each variable c_i^{-1}, (1.3) is satisfied on a set of points with a limit point in the domain of convergence. By analytic continuation, (1.3) is valid for all c_i, $1 \le i \le 4$, in the domain of convergence $|q^{-3}\prod_{i=1}^{2}(b_i/a_i)^2| < 1$. We also eliminate the assumption that $c_ic_j^{-1} \ne q^l$ by analytic continuation. This completes the proof of Theorem 1.1.

Remark 2.32. It should be mentioned that the analytic continuation argument used here is related to one used by Ismail [13] in an elementary proof of Ramanujan's ${}_1\psi_1$ summation theorem. Milne [18] gives a generalization of Ismail's argument for basic hypergeometric series in $U(n)$.

3. Proof of Theorem 1.7. In this section we will prove Theorem 1.7. We first prove the convergence condition $\mathrm{Re}(2\sum_{i=1}^{2}(\beta_i - \alpha_i)) > 3$ for the series in (1.9). Then we will prove (1.9) by taking the limit as $q \to 1$ of a terminating case of Theorem 1.1. An application of Carlson's theorem [6, p. 39] gives the general result.

To show the convergence of the series in (1.9), we follow an argument similar to the proof of the convergence of the series in (1.3). Given $t \in L$ such that $t = n_1\varphi_1 + n_2\varphi_2 + n_3\varphi_3$ with $n_1, n_2, n_3 \in \mathbb{Z}$, then there is a permutation σ of $\{1, 2, 3\}$ such that $t = l_1\varphi_{\sigma(1)} - l_2\varphi_{\sigma(3)}$, where $l_1 = n_{\sigma(1)} - n_{\sigma(2)}$ and $l_2 = n_{\sigma(2)} - n_{\sigma(3)}$ with $l_1, l_2 \ge 0$. As in

(2.3), it follows that

$$\prod_{i=1}^{2}\prod_{\substack{1\leqq j,k\leqq 3\\ j\neq k}}\frac{(\alpha_j+z_j-z_k)_{t_j-t_k}}{(\beta_i+z_j-z_k)_{t_j-t_k}}$$

$$=\prod_{i=1}^{2}\left\{\frac{(\alpha_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}(\alpha_i+z_{\sigma(2)}-z_{\sigma(1)})_{-l_1}}{(\beta_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}(\beta_i+z_{\sigma(2)}-z_{\sigma(1)})_{-l_1}}\right.$$

$$\cdot\frac{(\alpha_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}(\alpha_i+z_{\sigma(3)}-z_{\sigma(2)})_{-l_2}}{(\beta_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}(\beta_i+z_{\sigma(3)}-z_{\sigma(2)})_{-l_2}}$$

(3.1)

$$\left.\cdot\frac{(\alpha_i+z_{\sigma(1)}-z_{\sigma(3)})_{l_1+l_2}(\alpha_i+z_{\sigma(3)}-z_{\sigma(1)})_{-l_1-l_2}}{(\beta_i+z_{\sigma(1)}-z_{\sigma(3)})_{l_1+l_2}(\beta_i+z_{\sigma(3)}-z_{\sigma(1)})_{-l_1-l_2}}\right\}$$

$$=\prod_{i=1}^{2}\left\{\frac{(\alpha_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}(1-\beta_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}}{(\beta_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}(1-\alpha_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}}\right.$$

$$\cdot\frac{(\alpha_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}(1-\beta_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}}{(\beta_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}(1-\alpha_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}}$$

$$\left.\cdot\frac{(\alpha_i+z_{\sigma(1)}-z_{\sigma(3)})_{l_1+l_2}(1-\beta_i+z_{\sigma(1)}-z_{\sigma(3)})_{l_1+l_2}}{(\beta_i+z_{\sigma(1)}-z_{\sigma(3)})_{l_1+l_2}(1-\alpha_i+z_{\sigma(1)}-z_{\sigma(3)})_{l_1+l_2}}\right\}.$$

With t as above, also observe that

$$|(z_1+t_1-z_2-t_2)(-z_1-t_1+z_3+t_3)(-z_2-t_2+z_3+t_3)|$$

(3.2a)

$$\leqq k_1(l_1+1)(l_2+1)(l_1+l_2+1),$$

$$|(-2z_1-2t_1+z_2+t_2+z_3+t_3)(z_1+t_1-2z_2-2t_2+z_3+t_3)$$

(3.2b)

$$\cdot(-z_1-t_1-z_2-t_2+2z_3+2t_3)|\leqq k_2(l_1+l_2+1)^3$$

for some constants k_1, $k_2\geqq 0$.

From (3.1) and (3.2a, b) we see that the series in identity (1.9) converges absolutely if, for all $\sigma\in S_3$, the following series converges absolutely:

(3.5)

$$\sum_{l_1,l_2\geqq 0}(l_1+1)(l_2+1)(l_1+l_2+1)^4$$

$$\cdot\prod_{i=1}^{2}\left\{\frac{(\alpha_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}}{(\beta_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}}\right.$$

(3.3)

$$\cdot\frac{(1-\beta_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}(\alpha_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}(1-\beta_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}}{(1-\alpha_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}(\beta_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}(1-\alpha_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}}$$

$$\left.\cdot\frac{(\alpha_i+z_{\sigma(1)}-z_{\sigma(3)})_{l_1+l_2}(1-\beta_i+z_{\sigma(1)}-z_{\sigma(3)})_{l_1+l_2}}{(\beta_i+z_{\sigma(1)}-z_{\sigma(3)})_{l_1+l_2}(1-\alpha_i+z_{\sigma(1)}-z_{\sigma(3)})_{l_1+l_2}}\right\}.$$

Using the classical identity

$$\lim_{n\to\infty}\frac{n^X\Gamma(n)}{\Gamma(n+X)}=1,\tag{3.4}$$

then the series (3.3) converges absolutely if and only if the following series converges absolutely:

$$\sum_{l_1,l_2\geqq 0}\frac{(l_1+1)(l_2+1)(l_1+l_2+1)^4}{(l_1+l_2+1)^{2(\beta_1+\beta_2-\alpha_1-\alpha_2)}}\prod_{i=1}^{2}$$

$$\cdot\left\{\frac{(\alpha_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}}{(\beta_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}}\,\frac{(1-\beta_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}(\alpha_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}(1-\beta_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}}{(1-\alpha_i+z_{\sigma(1)}-z_{\sigma(2)})_{l_1}(\beta_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}(1-\alpha_i+z_{\sigma(2)}-z_{\sigma(3)})_{l_2}}\right\}.$$

Assuming that $\operatorname{Re}(2\sum_{i=1}^{2}(\beta_i-\alpha_i))>3$ and using (3.4) again, then the series (3.5) converges absolutely if the following series converges absolutely:

$$\left\{\sum_{l_1\geqq 0}\frac{(l_1+1)^2}{l_1^{2(\beta_1+\beta_2-\alpha_1-\alpha_2)}}\right\}\cdot\left\{\sum_{l_2\geqq 0}\frac{(l_2+1)^2}{l_2^{2(\beta_1+\beta_2-\alpha_1-\alpha_2)}}\right\}. \tag{3.6}$$

The convergence of (3.6) is implied by the assumption $\operatorname{Re}(2\sum_{i=1}^{2}(\beta_i-\alpha_i))>3$. Therefore under the same assumption the series in (1.9) converges absolutely.

We will now prove Theorem 1.7. The proof is quite similar to that of Theorems 8.3 and 8.9 of [10]. Set

$$\alpha_i=-\alpha_i'+z_i-z_{i+1}, \tag{3.7a}$$

$$\beta_i=1+\beta_i'+z_i-z_{i+1} \tag{3.7b}$$

for $i=1,2$. We will assume that

$$\operatorname{Re}(\alpha_i'),\ \operatorname{Re}(\beta_i')\geqq 0 \tag{3.8a}$$

for $i=1,2$ and

$$z_j\in\mathbb{R},\qquad 0\leqq z_j<\tfrac{1}{4} \tag{3.8b}$$

for $1\leqq j\leqq 4$. Then for a nonnegative integer r and $1\leqq j,\ k\leqq 3$, $j\neq k$, we have

$$\left|\frac{-\alpha_i'+z_i-z_{i+1}+z_j-z_k+r}{1+\alpha_i'-z_i+z_{i+1}+z_j-z_k+r}\right|<1 \tag{3.9}$$

for $i=1,2$ and a similar relation for β_i'. It follows that

$$\prod_{\substack{1\leqq j,k\leqq 3\\ j\neq k}}|(-\alpha_i'+z_i-z_{i+1}+z_j-z_k)_{t_j-t_k}|<1 \tag{3.10}$$

for $i=1,2$ and $t\in L$, with a similar relation for β_i'. Hence if we choose some α_i', $1\leqq i\leqq 2$, and fix all other parameters α_j', $j\neq i$, and β_1', β_2', z_1, z_2, and z_3, then we have

$$\operatorname{Re}(2(\beta_1+\beta_2-\alpha_j))=\operatorname{Re}(4+2(\beta_1'+\beta_2'+\alpha_j'))>3. \tag{3.11}$$

By a slight modification of the argument above showing convergence of the series in (1.9), it follows that the series in (1.9) is uniformly and absolutely convergent for $\operatorname{Re}(\alpha_i')\geqq 0$ and is regular and bounded in the variable α_i' for $\operatorname{Re}(\alpha_i')\geqq 0$.

Using Stirling's formula

$$\Gamma(a+s)=\sqrt{2\pi}\,s^{a+s-1/2}e^{-s}\left(1+O\left(\frac{1}{s}\right)\right)$$

as $s\to\infty$ in $S_\theta=\{s\colon|\arg s|<\theta\}$, $0<\theta<\pi$, and $a,\ s\in\mathbb{C}$, we can show that under conditions (3.8a, b) the right-hand side of (1.9) is regular and of polynomial growth in the region $\operatorname{Re}(\alpha_i')\geqq 0$ as a function of the variable α_i'. There is a similar statement if we choose the parameter β_i', $1\leqq i\leqq 2$, and fix the other parameters β_j', $j\neq i$, and α_1', α_2', z_1, z_2, and z_3. Subject to (3.8a, b), the series on the left-hand of (1.9) will be regular and bounded in the region $\operatorname{Re}(\beta_i')\geqq 0$ as a function of the variable β_i'. Also the right-hand side of (1.9), as a function of β_i', will be regular and of polynomial growth in this region.

Now set

$$a_j=q^{-k_j+z_j-z_{j+1}},\qquad b_j=q^{1+n_j+z_j-z_{j+1}}$$

for $j=1,2$ in (1.3), where the k_j and n_j are nonnegative integers. Then the left-hand side of (1.3) reduces to a finite sum of nonzero terms as the series terminates in both

directions in all summation indices. Similarly, the right-hand side of (1.3) reduces to a quotient of finite products. Taking the limit of (1.3) as $q \to 1$ with these assumptions, we obtain the special case of (1.9), where $\alpha_j' = k_j$ and $\beta_j' = n_j$ for $j = 1, 2$ and nonnegative integers k_j and n_j.

With assumptions as in (3.8a, b), we choose a parameter α_i', $1 \leqq i \leqq 2$, and fix the other parameters, subject to the conditions $\alpha_j' = k_j$ for $j \neq i$ and $\beta_j' = n_j$ for $j = 1, 2$, where the k_j and n_j are nonnegative integers. Let F be the difference between the left- and right-hand sides of (1.9). (The choice of sign is unimportant.) We view F as a function of α_i', as regular in the region $\mathrm{Re}(\alpha_i') \geqq 0$, and as vanishing when $\alpha_i' = 0, 1, 2, \cdots$. By Carlson's theorem [6, p. 39], F is identically zero as a function of α_i'. Repeating this procedure for the parameters α_j' and β_j' successively, we remove the restrictions that $\alpha_j' = k_j$ and $\beta_j' = n_j$ for nonnegative integers k_j and n_j. Finally, by analytic continuation the restrictions on the parameters z_j, $1 \leqq j \leqq l$, are removed. It follows that F vanishes identically as a function of all the parameters α_j', β_j', and z_j, when it is defined and converges. This completes the proof of Theorem 1.7.

REFERENCES

[1] G. E. ANDREWS, *Applications of basic hypergeometric functions*, SIAM Rev., 16 (1974), pp. 441-484.

[2] ———, *The Theory of Partitions*, Encyclopedia of Mathematics and Applications, Vol. 2, G.-C. Rota, ed., Addison-Wesley, Reading, MA, 1976.

[3] R. ASKEY AND J. A. WILSON, *A set of orthogonal polynomials that generalize the Racah coefficients of 6-j symbols*, SIAM J. Math. Anal., 10 (1979), pp. 1008-1016.

[4] ———, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 54 (1985), pp. 1-55.

[5] W. N. BAILEY, *Series of hypergeometric type which are infinite in both directions*, Quart. J. Math., 71 (1936), pp. 105-115.

[6] ———, *Generalized Hypergeometric Series*, Cambridge Math. Tract 32, Cambridge University Press, Cambridge, 1935. (Reprinted, Hafner, New York, 1964.)

[7] N. BOURBAKI, *Groupes et algebres de Lie*, Hermann, Paris, 1969, Chaps. 4-6.

[8] R. A. GUSTAFSON, *A Whipple's transformation for hypergeometric series in $U(n)$ and multivariable hypergeometric orthogonal polynomials*, SIAM J. Math. Anal., 18 (1987), pp. 495-530.

[9] ———, *Multilateral summation theorems for ordinary and basic hypergeometric series in $U(n)$*, SIAM J. Math. Anal., 18 (1987), pp. 1576-1596.

[10] ———, *The Macdonald identities for affine root systems of classical type and hypergeometric series very-well-poised on semisimple Lie algebras*, in Proc. Ramunujan Birth Centenary Year International Symposium on Analysis, N. K. Thakare, ed. Pune, India, December 1987.

[11] R. A. GUSTAFSON AND S. C. MILNE, *Schur functions, Good's identity, and hypergeometric series well poised in $U(n)$*, Adv. in Math., 48 (1983), pp. 177-188.

[12] J. E. HUMPHREYS, *Introduction to Lie Algebras and Representation Theory*, Springer-Verlag, Berlin, New York, 1972.

[13] M. E. H. ISMAIL, *A simple proof of Ramanujan's ${}_1\psi_1$ sum*, Proc. Amer. Math. Soc., 63 (1977), pp. 185-186.

[14] J. D. LOUCK AND L. C. BIEDENHARN, *Canonical unit adjoint tensor operators in $U(n)$*, J. Math. Phys., 11 (1970), pp. 2368-2414.

[15] I. G. MACDONALD, *Affine root systems and Dedekind's η-function*, Invent. Math., 15 (1972), pp. 91-143.

[16] ———, *Symmetric Functions and Hall Polynomials*, Oxford University Press, London, New York, 1979.

[17] S. C. MILNE, *An elementary proof of the Macdonald identities for $A_l^{(1)}$*, Adv. in Math., 57 (1985), pp. 34-70.

[18] ———, *A $U(n)$ generalization of Ramanujan's ${}_1\psi_1$ summation*, J. Math. Anal. Appl., 118 (1986), pp. 263-277.

[19] J. A. WILSON, *Some hypergeometric orthogonal polynomials*, SIAM J. Math. Anal., 11 (1980), pp. 690-701.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 523-535, March 1990
© 1990 Society for Industrial and Applied Mathematics
013

ERROR BOUNDS FOR A UNIFORM ASYMPTOTIC EXPANSION OF THE LEGENDRE FUNCTION $Q_n^{-m}(\cosh z)$ *

C. L. FRENZEN†

Abstract. For fixed m with $m+\frac{1}{2}>0$, an asymptotic expansion for large n is obtained for the Legendre function $Q_n^{-m}(\cosh z)$ that is uniformly valid for z in the unbounded interval $[0,\infty)$. This method is based on an integral representation of the function. The coefficients in the expansion satisfy a recurrence relation, and simple computable bounds are provided for the error terms associated with the expansion.

Key words. error bound, uniform asymptotic expansion, Legendre function

AMS(MOS) subject classifications. primary 41A60, 33A45

1. Introduction. Recently several authors have obtained asymptotic expansions of Legendre functions directly from their integral representations. The derivation of these expansions provides an alternative to Olver's differential equation approach, which gives asymptotic expansions of $P_n^{-m}(\cosh z)$ and $Q_n^m(\cosh z)$ for large n and fixed $m\geqq 0$ that are uniformly valid for z in a *complex* domain containing the interval $[0,\infty)$ [3, Chap. 12, §§ 12, 13]. Olver also provides numerical bounds for the error terms associated with these expansions.

Ursell [7] has given an alternative derivation of these expansions from integral representations of the Legendre functions, and has presented a new method for constructing error bounds. His bounds, though, are not computable and are valid only for bounded *complex* z. Starting from an integral representation Shivakumar and Wong [6] give an asymptotic expansion of $P_n^{-m}(\cosh z)$ for real z and fixed m, $m+\frac{1}{2}>0$, and construct error bounds comparable to those given by Olver, *computable* and uniformly valid for z in the infinite interval $[0,\infty)$. Although it is capable of deriving an asymptotic expansion for the other Legendre function $Q_n^{-m}(\cosh z)$, their method could not provide error bounds.

In this paper, we provide computable error bounds for an asymptotic expansion of $Q_n^{-m}(\cosh z)$ obtained from an integral representation. These bounds are uniformly valid for z in the infinite interval $[0,\infty)$. Our expansion differs from, but is equivalent to, those given by Olver [3, p. 465] and Ursell [7]. More precisely, we show that for $m+\frac{1}{2}>0$, $n-m+1>0$, and for any positive integer p

$$e^{mi\pi}Q_n^{-m}(\cosh z)=\left(\frac{z}{\sinh z}\right)^{1/2}\left[\sum_{\nu=0}^{p-1} d_\nu(z)\frac{K_{m+\nu}(uz)}{u^{m+\nu}}+\varepsilon_p(z,u)\right] \tag{1.1}$$

where $u=n+\frac{1}{2}$ and $K_\nu(z)$ is the modified Bessel function. The coefficients $d_\nu(z)$ are analytic functions of z, and can be obtained recursively. The first three are given by

$$d_0(z)=1,\qquad d_1(z)=\frac{\Gamma(m+\frac{3}{2})}{\Gamma(m+\frac{1}{2})}\left(m-\frac{1}{2}\right)\frac{z\coth z-1}{2z}, \tag{1.2}$$

$$d_2(z)=\frac{\Gamma(m+\frac{5}{2})}{\Gamma(m+\frac{3}{2})}\left\{\frac{1}{6}\left(m-\frac{1}{2}\right)\left[1+\frac{3}{z^2}-\frac{3}{z}\coth z\right]+\frac{1}{8}\left(m-\frac{1}{2}\right)\left(m-\frac{3}{2}\right)\left(\frac{(1-z\coth z)^2}{z^2}\right)\right\}.$$

The error term $\varepsilon_p(z,u)$ satisfies the inequality

$$|\varepsilon_p(z,u)|\leqq\frac{\Gamma(m+p+\frac{1}{2})}{\Gamma(m+\frac{1}{2})}\left(\frac{2z}{2+z}\right)^p M_p\frac{K_{m+p}(uz)}{u^{m+p}} \tag{1.3}$$

* Received by the editors September 6, 1988; accepted for publication March 23, 1989.

† Department of Mathematics, Southern Methodist University, Dallas, Texas 75275. Present address, Department of Mathematics, Naval Postgraduate School, Monterey, California 93943. This research was supported in part by National Science Foundation grant DMS-8718941.

for $-\frac{1}{2}<m\leqq\frac{1}{2}$, and

$$|\varepsilon_p(z,u)|\leqq\frac{\Gamma(m+p+\frac{1}{2})}{\Gamma(m+\frac{1}{2})}\left(\frac{2z}{2+z}\right)^p M_p\, e^{-(m-1/2)z}\,\frac{K_{m+p}([u-(m-\frac{1}{2})]z)}{[u-(m-\frac{1}{2})]^{m+p}} \tag{1.4}$$

for $m>\frac{1}{2}$, where $0\leqq z<\infty$ and $n>-\frac{1}{2}$. Note that since $n-m+1>0$, $u-(m-\frac{1}{2})>0$. M_p is a computable constant independent of u and z, and the first three are

$$\begin{aligned} M_1&=|m-\tfrac{1}{2}|, \qquad M_2=\frac{|m-\frac{1}{2}|}{2}+\frac{|m-\frac{1}{2}||m-\frac{3}{2}|}{2}\\ M_3&=\frac{|m-\frac{1}{2}|}{6}+\frac{|m-\frac{1}{2}||m-\frac{3}{2}|}{2}+\frac{|m-\frac{1}{2}||m-\frac{3}{2}||m-\frac{5}{2}|}{6}. \end{aligned} \tag{1.5}$$

Regarding the form of the bound in (1.4), later we show that for $m>\frac{1}{2}$,

$$1<e^{-(m-1/2)z}\,\frac{K_{p+m}([u-(m-\frac{1}{2})]z)}{K_{p+m}(uz)}\leqq\left[\frac{u}{u-(m-\frac{1}{2})}\right]^{p+m} \tag{1.6}$$

for z in $[0,\infty)$. Equation (1.6) implies that

$$e^{-(m-1/2)z}K_{p+m}([u-(m-\tfrac{1}{2})]z)=K_{p+m}(uz)(1+o(1)) \tag{1.7}$$

uniformly for z in $[0,\infty)$ as $u\to\infty$.

Expansion (1.1) can be rearranged to agree with those obtained by Olver and Ursell, but, in doing so, the simple estimate for the remainder given in (1.3) and (1.4) is lost. A comparison of our results with those of Olver and a discussion of our bounds is given in the final section of this paper.

2. Derivation of the expansion. For $m+\frac{1}{2}>0$ and $u-m+\frac{1}{2}>0$ we have the integral representation [1, § 3.7]

$$e^{mi\pi}Q_n^{-m}(\cosh z)=\left(\frac{\pi}{2}\right)^{1/2}\frac{(\sinh z)^{-m}}{\Gamma(m+\frac{1}{2})}\int_z^\infty e^{-ut}(\cosh t-\cosh z)^{m-1/2}\,dt \tag{2.1}$$

where $u=n+\frac{1}{2}$. As in [6], we first obtain a series expansion for the integrand in (2.1). Beginning with the result [8, Form. (3), p. 140]

$$\left(\frac{2}{\pi z}\right)^{1/2}\cos\sqrt{z^2-2z\theta}=\sum_{m=0}^{\infty}\frac{\theta^m}{m!}J_{m-1/2}(z) \tag{2.2}$$

valid for all complex θ and z, we replace z by iz and θ by $i\theta$ in (2.2) to obtain

$$\left(\frac{2}{\pi z}\right)^{1/2}\cosh\sqrt{z^2-2z\theta}=\sum_{m=0}^{\infty}\frac{(-1)^m}{m!}\theta^m I_{m-1/2}(z). \tag{2.3}$$

Here $J_\nu(z)$ and $I_\nu(z)$ are the Bessel functions of the first kind and the modified Bessel function, respectively, and $J_\nu(iz)=e^{\nu\pi i/2}I_\nu(z)$. Putting $z^2-2z\theta=t^2$ in (2.3) and recalling that

$$I_{-1/2}(z)=\left(\frac{2}{\pi z}\right)^{1/2}\cosh z, \qquad I_{1/2}(z)=\left(\frac{2}{\pi z}\right)^{1/2}\sinh z$$

then yields

$$\cosh t-\cosh z=\frac{\sinh z}{2z}(t^2-z^2)\sum_{\nu=0}^{\infty}\phi_\nu(z)(t^2-z^2)^\nu \tag{2.4}$$

where $\phi_0(z)=1$ and

$$\phi_\nu(z)=\frac{1}{2^\nu(\nu+1)!}\frac{I_{\nu+1/2}(z)}{z^\nu I_{1/2}(z)}, \qquad \nu\geqq 1. \tag{2.5}$$

Note from (2.5) and the asymptotic behavior of $I_\nu(z)$ as $z\to\infty$ (see [3, p. 251)] that

$$\phi_\nu(z)=O(z^{-\nu}) \qquad (z\to\infty). \tag{2.6}$$

The expression

$$(\cosh t-\cosh z)^{m-1/2}=\left(\frac{\sinh z}{2z}\right)^{m-1/2}(t^2-z^2)^{m-1/2}\sum_{\nu=0}^{\infty}\psi_\nu(z)(t^2-z^2)^\nu \tag{2.7}$$

now follows immediately for t sufficiently close to z, where the coefficients $\psi_\nu(z)$ satisfy the recurrence relation

$$\psi_{\nu+1}(z)=\sum_{j=0}^{\nu}\left[\left(m-\frac{1}{2}\right)-\frac{j}{\nu+1}\left(m+\frac{1}{2}\right)\right]\phi_{\nu+1-j}(z)\psi_j(z), \tag{2.8}$$

$\nu=0, 1, 2, \cdots$, with $\psi_0(z)=1$ (see [4]). As in [6],

$$\psi_1(z)=-\frac{1}{4}\left(m-\frac{1}{2}\right)\frac{1-z\coth z}{z^2}, \tag{2.9}$$

$$\begin{aligned}\psi_2(z)=&\frac{1}{24}\left(m-\frac{1}{2}\right)\left[\frac{1}{z^2}\left(1+\frac{3}{z^2}\right)-\frac{3}{z^3}\coth z\right]\\&+\frac{1}{32}\left(m-\frac{1}{2}\right)\left(m-\frac{3}{2}\right)\left(\frac{1-z\coth z}{z^2}\right)^2.\end{aligned} \tag{2.10}$$

Note from (2.6) and (2.8) that

$$\psi_\nu(z)=O(z^{-\nu}) \qquad (z\to\infty). \tag{2.11}$$

We now introduce the remainder $\Delta_p(z, t)$ defined by

$$\left(\frac{\cosh t-\cosh z}{t^2-z^2}\frac{2z}{\sinh z}\right)^{m-1/2}=\sum_{\nu=0}^{p-1}\psi_\nu(z)(t^2-z^2)^\nu+(t^2-z^2)^p\Delta_p(z, t), \qquad t\geqq z. \tag{2.12}$$

Inserting (2.12) into (2.1) and using the formula (see [8, § 6.3, p. 185])

$$K_\nu(uz)=\frac{\pi^{1/2}(2z)^{-\nu}u^\nu}{\Gamma(\nu+\frac{1}{2})}\int_z^\infty e^{-ut}(t^2-z^2)^{\nu-1/2}\,dt, \tag{2.13}$$

then yields

$$e^{mi\pi}Q_n^{-m}(\cosh z)=\left(\frac{z}{\sinh z}\right)^{1/2}\left[\sum_{\nu=0}^{p-1}d_\nu(z)\frac{K_{m+\nu}(uz)}{u^{m+\nu}}+\varepsilon_p(z, u)\right] \tag{2.14}$$

where

$$\varepsilon_p(z, u)=\frac{\pi^{1/2}}{(2z)^m\Gamma(m+\frac{1}{2})}\int_z^\infty e^{-ut}(t^2-z^2)^{p+m-1/2}\Delta_p(z, t)\,dt. \tag{2.15}$$

The coefficient $d_\nu(z)$ is given explicitly by

$$d_\nu(z)=\frac{\Gamma(m+\nu+\frac{1}{2})}{\Gamma(m+\frac{1}{2})}(2z)^\nu\psi_\nu(z) \tag{2.16}$$

(see (1.2)). Note from (2.1) and (2.12) that when $m=\frac{1}{2}$, $d_j(z)$ and $\varepsilon_j(z,u)$ are both zero for all $j\geqq 1$.

In the following sections we will show that, for $-\frac{1}{2}<m\leqq\frac{1}{2}$,

$$|\Delta_p(z,t)|\leqq\frac{M_p}{(2+z)^p},\qquad t\geqq z, \tag{2.17}$$

while for $m>\frac{1}{2}$

$$|\Delta_p(z,t)|\leqq\frac{M_p}{(2+z)^p}\,e^{(t-z)(m-1/2)},\qquad t\geqq z \tag{2.18}$$

where M_p is a computable constant independent of u and z explicitly given in (4.10). The first three are given in (1.5), the desired estimates in (1.3) and (1.4) now follow from (2.13), (2.15), (2.17), and (2.18).

3. Some preliminary results. To establish the estimates in (2.17) and (2.18) we will study the function

$$\sigma(z,t)=\frac{2z}{\sinh z}\left(\frac{\cosh t-\cosh z}{t^2-z^2}\right) \tag{3.1}$$

for $t\geqq z$, with $\sigma(z,z)=1$.

LEMMA 1. *For $t\geqq z$,*

$$\begin{aligned}&(\sigma(z,t))^{m-1/2}\leqq 1,\qquad -\tfrac{1}{2}<m\leqq\tfrac{1}{2},\\ &(\sigma(z,t))^{m-1/2}\leqq e^{(t-z)(m-1/2)},\qquad m>\tfrac{1}{2}.\end{aligned} \tag{3.2}$$

Proof. From the identity

$$\frac{\cosh t-\cosh z}{t^2-z^2}=\frac{1}{2}\,\frac{\sinh\left(\frac{t+z}{2}\right)}{\left(\frac{t+z}{2}\right)}\cdot\frac{\sinh\left(\frac{t-z}{2}\right)}{\left(\frac{t-z}{2}\right)}, \tag{3.3}$$

it follows that

$$\sigma(z,t)=\left(\frac{z}{\sinh z}\,\frac{\sinh\left(\frac{t+z}{2}\right)}{\left(\frac{t+z}{2}\right)}\right)\left(\frac{\sinh\left(\frac{t-z}{2}\right)}{\left(\frac{t-z}{2}\right)}\right). \tag{3.4}$$

Each term in parentheses in (3.4) is greater than or equal to one for $t\geqq z$, yielding the first inequality in (3.2). For the second, we again use (3.3) to conclude

$$\sigma(z,t)=(e^t)\left(\frac{z}{\sinh z}\right)\left(\frac{1-e^{-(t+z)}}{t+z}\right)\left(\frac{1-e^{-(t-z)}}{t-z}\right). \tag{3.5}$$

Since the function $(1-e^{-\theta})/\theta$ decreases monotonically from one for θ in $[0,\infty)$, for $t\geqq z$ (3.5) implies

$$\sigma(z,t)\leqq(e^t)\left(\frac{z}{\sinh z}\right)\left(\frac{1-e^{-2z}}{2z}\right)=e^{t-z}, \tag{3.6}$$

and this gives the second inequality in (3.2). □

From (2.4) it follows that σ is a function of t^2-z^2. Denoting this function by G and putting $x=t^2-z^2$ with $x\geqq 0$ for $t\geqq z$, we have

$$\begin{aligned}\sigma(z,t)=G(x)&\equiv\sum_{\nu=0}^{\infty}\phi_\nu(z)x^\nu\\&=\frac{2z}{\sinh z}\,\frac{\cosh\sqrt{z^2+x}-\cosh z}{x}.\end{aligned}\tag{3.7}$$

In what follows it will be convenient to *temporarily* replace z by $\sqrt{z}$ in studying G. Consequently, x becomes t^2-z, where $t\geqq\sqrt{z}$. After we have obtained the required results we will again replace $\sqrt{z}$ by z. So, from (3.7) we will study the function

$$G(x)=\frac{2\sqrt{z}}{\sinh\sqrt{z}}\,\frac{\cosh\sqrt{z+x}-\cosh\sqrt{z}}{x}=\sum_{\nu=0}^{\infty}\phi_\nu(\sqrt{z})x^\nu,\qquad x\geqq 0.\tag{3.8}$$

Equation (3.8) implies that

$$G(x)=\frac{2\sqrt{z}}{\sinh\sqrt{z}}\sum_{k=0}^{\infty}h^{(k+1)}(z)\frac{x^k}{(k+1)!}\tag{3.9}$$

where $h(z)=\cosh\sqrt{z}$ and the Maclaurin series expansion of $\cosh\sqrt{z+x}$ has been used.

LEMMA 2. *With* $h(z)=\cosh\sqrt{z}$, $r=0,1,2,\cdots$, *and* $z\geqq 0$,

$$h^{(r+1)}(z)h^{(r+1)}(z)>h^{(r)}(z)h^{(r+2)}(z).\tag{3.10}$$

Proof. With

$$h(z)=\cosh\sqrt{z}=\sum_{k=0}^{\infty}\frac{z^k}{(2k)!}\tag{3.11}$$

and

$$\frac{d^r}{dz^r}z^k=\frac{\Gamma(k+1)}{\Gamma(k-r+1)}z^{k-r},\tag{3.12}$$

we have

$$\begin{aligned}h^{(r)}(z)&=\sum_{k=r}^{\infty}\frac{\Gamma(k+1)}{\Gamma(k-r+1)}\,\frac{z^{k-r}}{\Gamma(2k+1)}\\&=\sum_{j=0}^{\infty}\frac{\Gamma(r+j+1)}{\Gamma(j+1)}\,\frac{z^j}{\Gamma(2(r+j)+1)}\\&=\frac{\pi^{1/2}}{2^{2r}}\sum_{j=0}^{\infty}\frac{z^j}{2^{2j}j!\Gamma(r+j+\frac{1}{2})}\end{aligned}\tag{3.13}$$

where the duplication formula for the gamma function has been used in the last equality in (3.13). From the latter equation it follows that $h^{(r)}(0)>0$, while

$$\begin{vmatrix}h^{(r)}(0) & h^{(r+1)}(0)\\ h^{(r+1)}(0) & h^{(r+2)}(0)\end{vmatrix}=\frac{\pi}{2^{4r+4}\Gamma(r+\frac{1}{2})\Gamma(r+\frac{3}{2})}\left[\frac{1}{r+\frac{3}{2}}-\frac{1}{r+\frac{1}{2}}\right]<0\tag{3.14}$$

for $r=0,1,2,\cdots$, where the vertical lines in (3.14) indicate a determinant. These results can be combined by writing

$$\varepsilon_k\begin{vmatrix}h^{(r)}(0)&\cdots&h^{(r+k)}(0)\\ h^{(r+k)}(0)&\cdots&h^{(r+2k)}(0)\end{vmatrix}>0,\quad k=0,1,\quad r=0,1,2,3,\cdots\tag{3.15}$$

where $\varepsilon_0=1$, $\varepsilon_1=-1$. We now employ the following result from Theorem 8.4 of [2, p. 84].

LEMMA. *Suppose $f(z)$ has a power series expansion about the origin that is convergent for $|z|<\rho$. If $f(z)$ satisfies*

$$\varepsilon_k\begin{vmatrix} f^{(r)}(0) & f^{(r+1)}(0) & \cdots & f^{(r+k)}(0) \\ f^{(r+1)}(0) & & \cdots & f^{(r+k+1)}(0) \\ \vdots & & & \vdots \\ f^{(r+k)}(0) & & \cdots & f^{(r+2k)}(0) \end{vmatrix}>0$$

for $r=0,1,2,\cdots$, $k=0,1,2,\cdots,n$, where $\varepsilon_0,\varepsilon_1,\cdots,\varepsilon_n$ is a prescribed sequence of plus or minus signs, then

$$\varepsilon_k\begin{vmatrix} f^{(r)}(z) & f^{(r+1)}(z) & \cdots & f^{(r+k)}(z) \\ f^{(r+1)}(z) & & \cdots & f^{(r+k+1)}(z) \\ \vdots & & & \vdots \\ f^{(r+k)}(z) & & \cdots & f^{(r+2k)}(z) \end{vmatrix}>0$$

for $0\leqq z<\rho$, $k=0,1,2,\cdots,n$, $r=0,1,2,\cdots$.

With $\rho=\infty$, $n=1$, $f(z)=h(z)=\cosh\sqrt{z}$, $\varepsilon_0=1$, $\varepsilon_1=-1$, the above lemma and (3.15) allow us to conclude that

$$-\begin{vmatrix} h^{(r)}(z) & h^{(r+1)}(z) \\ h^{(r+1)}(z) & h^{(r+2)}(z) \end{vmatrix}>0 \tag{3.16}$$

for $z\geqq 0$, $r=0,1,2,\cdots$, and this is (3.10). □

Inequality (3.10) implies

$$\frac{h^{(r+2)}(z)}{h^{(r+1)}(z)}<\frac{h^{(r+1)}(z)}{h^{(r)}(z)}$$

for $z\geqq 0$, $r=0,1,2,\cdots$, so that

$$\frac{h^{(r+k+1)}(z)}{h^{(r+k)}(z)}<\frac{h^{(r+k)}(z)}{h^{(r+k-1)}(z)}<\cdots<\frac{h^{(l+1)}(z)}{h^{(l)}(z)}<\cdots<\frac{h^{(2)}(z)}{h^{(1)}(z)} \tag{3.17}$$

for $z\geqq 0$, $r>0$, and $k\geqq 1$. In (3.17) the product of the first k terms on the left (setting $l=r+1$) is strictly less than the product of the last k terms on the right (setting $l=k$). Canceling numerators and denominators and admitting the possibility that $r=0$ or $k=0$ then implies

$$\frac{h^{(r+k+1)}(z)}{h^{(r+1)}(z)}\leqq\frac{h^{(k+1)}(z)}{h^{(1)}(z)} \tag{3.18}$$

for $z\geqq 0$, $r\geqq 0$, and $k\geqq 0$.

Let

$$A_r(z)=(r+1)\phi_r(\sqrt{z}) \tag{3.19}$$

where ϕ_r is defined in (2.5). The inequality in (3.18) now gives the following result for G defined in (3.8).

LEMMA 3. *For all $z>0$, $\xi\geqq 0$ and $r=0,1,2,\cdots$*

$$\frac{G^{(r)}(\xi)}{r!\,G(\xi)}\leqq A_r(z) \tag{3.20}$$

where $A_r(z)$ is given in (3.19).

Proof. From (3.9) and (3.12),

$$G^{(r)}(\xi)=\sum_{k=0}^{\infty}\frac{h^{(r+k+1)}(z)}{r+k+1}\frac{\xi^k}{k!}, \tag{3.21}$$

while from (3.8) and (3.9) it follows that, since $h^{(1)}(z)=\sinh\sqrt{z}/2\sqrt{z}$,

$$\frac{h^{(r+1)}(z)}{h^{(1)}(z)}=(r+1)!\,\phi_r(\sqrt{z})=r!\,A_r(z). \tag{3.22}$$

Therefore

$$\frac{G^{(r)}(\xi)}{r!\,G(\xi)}=A_r(z)\frac{\sum_{k=0}^{\infty}\frac{1}{r+k+1}\frac{h^{(r+k+1)}(z)}{h^{(r+1)}(z)}\frac{\xi^k}{k!}}{\sum_{k=0}^{\infty}\frac{1}{k+1}\frac{h^{(k+1)}(z)}{h^{(1)}(z)}\frac{\xi^k}{k!}}. \tag{3.23}$$

By (3.18) the series in the denominator on the right-hand side of (3.23) is a majorant of the series in the numerator, and this gives (3.20). □

Finally, we establish some upper bounds for $A_r(z)$. From (3.19) and (2.5)

$$A_r(z)=(r+1)\phi_r(\sqrt{z})=\frac{I_{r+1/2}(\sqrt{z})}{r!(2\sqrt{z})^r I_{1/2}(\sqrt{z})}. \tag{3.24}$$

Because $I_{r+1/2}(\sqrt{z})\leqq I_{1/2}(\sqrt{z})$ (see [3, p. 251]), it follows from (3.24) that

$$A_r(z)\leqq\frac{1}{r!(2\sqrt{z})^r}. \tag{3.25}$$

From the power series definition of $I_\nu(z)$ [3, p. 60] and the result

$$\frac{\Gamma(r+1)}{\Gamma(r+s+\frac{3}{2})}\leqq\frac{1}{\Gamma(s+\frac{3}{2})},$$

for $r=0,1,2,\cdots$, $s=0,1,2,\cdots$, an argument similar to that for (3.23) applied to (3.24) yields the following inequality:

$$A_r(z)\leqq\frac{1}{2^{2r}(r!)^2}. \tag{3.26}$$

For $0\leqq\sqrt{z}\leqq 2$, $2+\sqrt{z}\leqq 4$ and (3.26) implies

$$A_r(z)\leqq\frac{1}{(2+\sqrt{z})^r(r!)^2}. \tag{3.27}$$

For $\sqrt{z}\geqq 2$, $2\sqrt{z}\geqq 2+\sqrt{z}$ and (3.25) implies

$$A_r(z)\leqq\frac{1}{(2+\sqrt{z})^r r!}. \tag{3.28}$$

Together, (3.27) and (3.28) give

$$A_r(z)\leqq\frac{1}{(2+\sqrt{z})^r r!}, \tag{3.29}$$

for all $z\geqq 0$, $r=0,1,2,\cdots$.

4. Proof of (2.17) and (2.18). Changing z to $\sqrt{z}$ in (2.12) and recalling the definition of G in (3.8) gives

$$(G(x))^{m-1/2}=\sum_{\nu=0}^{p-1}\psi_\nu(\sqrt{z})x^\nu+x^p r_p(\sqrt{z},x),\qquad x\geqq 0 \tag{4.1}$$

where

$$r_p(\sqrt{z},x)=\Delta_p(\sqrt{z},\sqrt{x+z}). \tag{4.2}$$

We now use di Bruno's formula for the nth derivative of a composite function (see, for example, [5, p. 34]):

$$\frac{d^p}{dx^p}f(g(x))=\sum_{k=1}^{p}f^{(k)}(g(x))A_{pk}(g^{(1)}(x),g^{(2)}(x),\cdots,g^{(p)}(x)),\qquad p\geqq 1 \tag{4.3}$$

where $f^{(k)}$ denotes the kth derivative of f and the coefficients A_{pk} (which do not depend on f) are given by the formula

$$A_{pk}(g^{(1)}(x),\cdots,g^{(p)}(x))=\sum\frac{p!}{m_1!m_2!\cdots m_p!}\left(\frac{g^{(1)}(x)}{1!}\right)^{m_1}\cdots\left(\frac{g^{(p)}(x)}{p!}\right)^{m_p}. \tag{4.4}$$

The summation in (4.4) extends over all nonnegative integers $m_1, m_2, \cdots, m_p$ satisfying

$$m_1+m_2+\cdots+m_p=k,\qquad m_1+2m_2+\cdots+pm_p=p. \tag{4.5}$$

Taylor's theorem applied to (4.1) now implies that there exists a ξ between zero and x such that

$$\begin{aligned} r_p(\sqrt{z},x)&=\frac{1}{p!}\frac{d^p}{ds^p}((G(s))^{m-1/2})\Big|_{s=\xi}\\ &=\frac{1}{p!}\sum_{k=1}^{p}\frac{\Gamma(m+\frac{1}{2})}{\Gamma(m+\frac{1}{2}-k)}(G(\xi))^{m-1/2-k}A_{pk}(G^{(1)}(\xi),\cdots,G^{(p)}(\xi)) \end{aligned} \tag{4.6}$$

where (4.3) and (3.12) have been employed. From (4.4) and (4.5), (4.6) can be written as

$$\begin{aligned} r_p(\sqrt{z},x)&=(G(\xi))^{m-1/2}\sum_{k=1}^{p}\\ &\cdot\left(\frac{\Gamma(m+\frac{1}{2})}{\Gamma(m+\frac{1}{2}-k)}\sum\frac{\left(\frac{G^{(1)}(\xi)}{1!G(\xi)}\right)^{m_1}\left(\frac{G^{(2)}(\xi)}{2!G(\xi)}\right)^{m_2}\cdots\left(\frac{G^{(p)}(\xi)}{p!G(\xi)}\right)^{m_p}}{m_1!m_2!\cdots m_p!}\right) \end{aligned} \tag{4.7}$$

where, for each k, the inner sum in (4.7) sums over the solutions of (4.5). From (3.20) and (3.29)

$$\left(\frac{G^{(r)}(\xi)}{r!G(\xi)}\right)^{m_r}\leqq(A_r(z))^{m_r}\leqq\frac{1}{(r!)^{m_r}}\frac{1}{(2+\sqrt{z})^{rm_r}}. \tag{4.8}$$

Using (4.8) and (4.5) in the inner sum in (4.7) and taking absolute values then implies

$$|r_p(\sqrt{z},x)|\leqq\frac{(G(\xi))^{m-1/2}}{(2+\sqrt{z})^p}M_p \tag{4.9}$$

where

$$M_p = \sum_{k=1}^{p} \left(\left| \frac{\Gamma(m+\frac{1}{2})}{\Gamma(m+\frac{1}{2}-k)} \right| \sum \frac{\left(\frac{1}{1!}\right)^{m_1}\left(\frac{1}{2!}\right)^{m_2}\cdots\left(\frac{1}{p!}\right)^{m_p}}{m_1!\,m_2!\cdots m_p!} \right), \tag{4.10}$$

with, as before, the inner sum satisfying (4.5). The first three M_p's are given in (1.5). Noting that $0 \leqq \xi \leqq x$ in (4.9) and that $x = t^2 - z$, there exists some t_1, $\sqrt{z} \leqq t_1 \leqq t$, so that $\xi = t_1^2 - z$. With (3.7) and (4.2), (4.9) becomes

$$|\Delta_p(\sqrt{z}, t)| \leqq \frac{(\sigma(\sqrt{z}, t_1))^{m-1/2}}{(2+\sqrt{z})^p} M_p. \tag{4.11}$$

Now replace $\sqrt{z}$ by z. If $-\frac{1}{2} < m \leqq \frac{1}{2}$, then (3.2) in Lemma 1 gives (2.17). On the other hand, if $m > \frac{1}{2}$, the fact that $t \geqq t_1$ implies

$$\sigma(z, t_1) = G(t_1^2 - z^2) \leqq G(t^2 - z^2) = \sigma(z, t), \tag{4.12}$$

since G is an increasing function by (3.8). Lemma 1, (4.11), and (4.12) then yield (2.18).

5. Proof of (1.6) and comparison with Olver's result. From [8, p. 206], with $x > 0$,

$$K_\nu(x) = \left(\frac{\pi}{2x}\right)^{1/2} \frac{e^{-x}}{\Gamma(\nu+\frac{1}{2})} \int_0^\infty e^{-t} t^{\nu-1/2} \left(1+\frac{t}{2x}\right)^{\nu-1/2} dt, \tag{5.1}$$

so that, for $z > 0$,

$$K_{p+m}(uz) = \left(\frac{\pi}{2uz}\right)^{1/2} \frac{e^{-uz}}{\Gamma(p+m+\frac{1}{2})} \int_0^\infty e^{-t} t^{p+m-1/2} \left(1+\frac{t}{2uz}\right)^{p+m-1/2} dt, \tag{5.2}$$

while

$$e^{-(m-1/2)z} K_{p+m}\left(\left[u-\left(m-\frac{1}{2}\right)\right]z\right) = \left(\frac{\pi}{2[u-(m-\frac{1}{2})]z}\right)^{1/2} \frac{e^{-uz}}{\Gamma(p+m+\frac{1}{2})} \cdot \int_0^\infty e^{-t} t^{p+m-1/2} \left(1+\frac{t}{2[u-(m-\frac{1}{2})]z}\right)^{p+m-1/2} dt. \tag{5.3}$$

From (5.2) and (5.3) it follows that for $m > \frac{1}{2}$

$$1 < e^{-(m-1/2)z} \frac{K_{p+m}([u-(m-\frac{1}{2})]z)}{K_{p+m}(uz)}. \tag{5.4}$$

Since, for $\nu > 0$,

$$K_\nu(z) \sim \tfrac{1}{2}\Gamma(\nu)(\tfrac{1}{2}z)^{-\nu} \qquad (z \to 0) \tag{5.5}$$

[3, p. 252], (5.4) holds for z in $[0, \infty)$. Multiplying both sides of (5.2) by $(uz)^{p+m}$ and both sides of (5.3) by $[u-(m-\frac{1}{2})]^{p+m}$ implies

$$(uz)^{p+m} K_{p+m}(uz) > ([u-(m-\tfrac{1}{2})]z)^{p+m}\, e^{-(m-1/2)z} K_{p+m}([u-(m-\tfrac{1}{2})]z), \tag{5.6}$$

or

$$\frac{e^{-(m-1/2)z} K_{p+m}([u-(m-\frac{1}{2})]z)}{K_{p+m}(uz)} < \left(\frac{u}{u-(m-\frac{1}{2})}\right)^{p+m}. \tag{5.7}$$

As $z \to 0^+$, the inequality in (5.7) becomes an equality (see (5.5)). When we take this possibility into account, (5.4) and (5.7) give (1.6) for $m > \frac{1}{2}$ and z in $[0, \infty)$.

It is difficult to compare our result to earlier ones. Ursell's bounds [7] break down when z becomes unbounded. Olver [3, p. 465] provides a uniform asymptotic expansion with error bounds for $Q_n^m(\cosh z)$ for large n, fixed $m \geqq 0$, and z in a complex domain containing $[0, \infty)$, whereas we have provided a uniform asymptotic expansion with error bounds for $Q_n^{-m}(\cosh z)$ for large n, fixed $m+\frac{1}{2}>0$, and z in $[0, \infty)$. Although the connection formula (see [3, p. 178])

$$Q_n^{-m}(\cosh z) = e^{-2m\pi i}\frac{\Gamma(n-m+1)}{\Gamma(n+m+1)}Q_n^m(\cosh z) \tag{5.8}$$

relates $Q_n^{-m}(\cosh z)$ and $Q_n^m(\cosh z)$ the presence of the gamma function ratio makes it difficult to compare our bound with Olver's. Additionally, Olver's expansion involves only two Besssel functions, K_m and K_{m+1}, while our expansion involves all the Bessel functions K_{m+p}, $p = 0, 1, 2, \cdots$. Our expansion can be rearranged to agree with Olver's, but the simplicity of the error estimates (1.3), (1.4) is then lost. As in [6], our coefficients $d_\nu(z)$ (given in (2.16)) are simple to calculate, and because $d_\nu(z) = O(z^\nu)$, $(z \to 0)$, our expansion is useful for small z.

Perhaps the best we can do to compare the two expansions is to take the one-term approximation that each gives when, in our expansion, $-m \geqq 0$. This occurs when $-\frac{1}{2} < m \leqq 0$, so if we substitute $-m$ into Olver's expansion, with $0 \leqq -m < \frac{1}{2}$, we can compare the two (see [6] for a similar comparison). In this case, the one-term approximation obtained from our expansion is (see (1.1) and (1.3) with $p = 1$)

$$e^{mi\pi}Q_n^{-m}(\cosh z) = \left(\frac{z}{\sinh z}\right)^{1/2}\left[\frac{K_m(uz)}{u^m} + \varepsilon_1(z, u)\right] \tag{5.9}$$

where

$$|\varepsilon_1(z, u)| \leqq \frac{2z}{2+z}\left|m^2 - \frac{1}{4}\right|\frac{K_{m+1}(uz)}{u^{m+1}}. \tag{5.10}$$

Substituting $-m \geqq 0$ into the one-term approximation obtained from Olver's expansion gives (see [3, eqs. (12.13)-(12.15), with $p = 0$, p. 465])

$$e^{m\pi i}Q_n^{-m}(\cosh z) = \frac{1}{1+\delta_1}\left(\frac{z}{\sinh z}\right)^{1/2}\left[\frac{K_m(uz)}{u^m} + \frac{\eta_{1,2}(u, z^2)}{u^m}\right] \tag{5.11}$$

where we have used the fact that $K_{-m}(uz) = K_m(uz)$ [3, p. 252]. In (5.11) δ_1 is a constant satisfying

$$|\delta_1| \leqq \lambda_1(-m)\frac{|m^2-\frac{1}{4}|}{2u}\exp\left[\lambda_1(-m)\frac{|m^2-\frac{1}{4}|}{2u}\right] \tag{5.12}$$

and

$$\frac{|\eta_{1,2}(u, z^2)|}{u^m} \leqq \lambda_1(-m)K_m(uz)\exp\left[\frac{\lambda_1(-m)}{u}V_{z,\infty}\{zB_0^{-m}(z^2)\}\right]\frac{V_{z,\infty}\{zB_0^{-m}(z^2)\}}{u^{m+1}} \tag{5.13}$$

where

$$\lambda_1(-m) = \sup_{x\in(0,\infty)}\{2xI_{-m}(x)K_m(x)\}, \tag{5.14}$$

$$zB_0^{-m}(z^2) = \frac{-(m^2-\frac{1}{4})}{2}\left(\coth z - \frac{1}{z}\right), \tag{5.15}$$

and

$$V_{z,\infty}\{zB_0^{-m}(z^2)\} = \frac{|m^2-\frac{1}{4}|}{2}\int_z^\infty \left(\frac{1}{t^2}-\frac{1}{(\sinh t)^2}\right) dt. \tag{5.16}$$

Note that

$$V_{0,\infty}\{zB_0^{-m}(z^2)\} = \frac{|m^2-\frac{1}{4}|}{2}. \tag{5.17}$$

Olver assumes that $u = n+\frac{1}{2}$ is large enough so that $|\delta_1|<1$. To ensure this, from (5.12) it is sufficient to require that

$$\lambda_1(-m)\frac{|m^2-\frac{1}{4}|}{2u} < 0.567. \tag{5.18}$$

Since $V_{z,\infty}\{zB_0^{-m}(z^2)\} \leqq V_{0,\infty}\{zB_0^{-m}(z^2)\}$, from (5.17) and (5.18) it follows that

$$\exp\left[\frac{\lambda_1(-m)}{u} V_{z,\infty}\{zB_0^{-m}(z^2)\}\right] < 2. \tag{5.19}$$

With (5.19), (5.13) may be written as

$$\frac{|\eta_{1,2}(u,z^2)|}{u^m} \leqq 2\lambda_1(-m)K_m(uz)\frac{V_{z,\infty}\{zB_0^{-m}(z^2)\}}{u^{m+1}}. \tag{5.20}$$

Because of the presence of the variational operator (5.16) in Olver's bound (5.20), and the presence of the factor $2z/2+z$ in our bound (5.10), Olver's bound is more effective for large z while our bound is more effective for small z. To see this, note that for fixed m with $-\frac{1}{2}<m\leqq 0$, $\lambda_1(m)$ and $(1+\delta_1)^{-1}$ are constants. First, suppose $uz>1$. From [3, p. 250],

$$K_{m+1}(uz) = O\left(\left(\frac{\pi}{2uz}\right)^{1/2} e^{-uz}\right), \qquad K_m(uz) = O\left(\left(\frac{\pi}{2uz}\right)^{1/2} e^{-uz}\right). \tag{5.21}$$

Since $|m^2-\frac{1}{4}|>0$, (5.16) and (5.17) imply that

$$\frac{2z}{2+z} = o(V_{z,\infty}\{zB_0^{-m}(z^2)\}) \qquad (z\to 0^+). \tag{5.22}$$

Consequently, for small enough z, the bound in (5.10) is better than the bound in (5.20). However, when z is large, $V_{z,\infty}\{zB_0^{-m}(z^2)\} = O(1/z)$ while $2z/(2+z) = O(1)$ and the situation is reversed. Now suppose that $uz\leqq 1$. If z is small, from (5.5) with $(m+1)>0$,

$$\frac{2z}{2+z} K_{m+1}(uz) = O\left(\frac{2^m\Gamma(m+1)}{u^{m+1}z^m}\right). \tag{5.23}$$

For small z the variational operator (5.16) is $O(1)$, and for $-m>0$ (5.5) again gives

$$V_{z,\infty}\{zB_0^{-m}(z^2)\}K_m(uz) = O\left(\frac{2^{-m-1}\Gamma(-m)}{(uz)^{-m}}\right), \tag{5.24}$$

while if $-m=0$ (from [3, p. 252]),

$$V_{z,\infty}\{zB_0^{-m}(z^2)\}K_m(uz) = O(-\ln(uz)). \tag{5.25}$$

Because $m+1>-m$ (since $m>-\frac{1}{2}$), and $-m\geqq m$ (since $m\leqq 0$), comparing (5.23), (5.24), and (5.25) reveals that for small z the bound in (5.10) is more effective than

the bound in (5.20). On the other hand, if z is large and $uz \leqq 1$ (which is possible since $u = n + \frac{1}{2} > 0$ is all we require), then

$$\frac{2z}{2+z} K_{m+1}(uz) = O\left(\frac{2^m \Gamma(m+1)}{(uz)^{m+1}}\right). \tag{5.26}$$

When $-m > 0$,

$$V_{z,\infty}\{zB_0^{-m}(z^2)\} K_m(uz) = O\left(\frac{2^{-m-1}\Gamma(-m)}{(uz)^{-m} z}\right), \tag{5.27}$$

while if $-m = 0$, then

$$V_{z,\infty}\{zB_0^{-m}(z^2)\} K_m(uz) = O\left(\frac{-\ln(uz)}{z}\right). \tag{5.28}$$

Since $uz \leqq 1$, $(m+1) > -m$ and z is large, (5.26)-(5.28) indicate that for large z the bound in (5.20) is better than the bound in (5.10).

To conclude, as in [6] we can assess the bounds in (1.3) and (1.4) for all m by examining the ratio $R_p^{-m}(z, u)$ of these bounds to the first neglected term $d_p(z) K_{m+p}(uz)/u^{m+p}$. If $m = \frac{1}{2}$ the expansion truncates after the first term and all the rest are zero; see the remark following (2.16). If $-\frac{1}{2} < m < \frac{1}{2}$, using (1.3) and canceling a common factor gives

$$R_p^{-m}(z, u) = \frac{M_p (2z)^p/(2+z)^p}{(2z)^p \psi_p(z)} \tag{5.29}$$

where $\psi_p(z)$ is defined recursively in (2.8), $\psi_1(z)$ and $\psi_2(z)$ being given in (2.9) and (2.10). The above ratio is interesting because it leads to a comparison between the maximum absolute value of the error bound for z in $[0, \infty)$ and the maximum absolute value of the first neglected term for z in $[0, \infty)$; bounded z poses no problems. For any specific m satisfying $-\frac{1}{2} < m < \frac{1}{2}$ the maximum value of $(2z)^p |\psi_p(z)|$ in $[0, \infty)$ can be computed numerically. Since $\psi_p(z)$ is analytic in $[0, \infty)$ and $\psi_p(z) = O(z^{-p})$ as $z \to \infty$, this value is a finite nonzero number. On the other hand, the maximum value of $(2z)^p/(2+z)^p$ in $[0, \infty)$ is 2^p. Letting $\tilde{R}_p^{-m}$ denote the ratio of the maxima of the absolute values of the numerator and the denominator in (5.29) over $[0, \infty)$, it follows that $\tilde{R}_p^{-m}$ is finite for each $p \geqq 1$. It is also evident from (5.29) that the error bound in (1.3) mimics the behavior of the first neglected term near both $z = 0$ and $z = \infty$.

For $m > \frac{1}{2}$, the ratio of the bound in (1.4) to the first neglected term is

$$\begin{aligned} R_p^{-m}(z, u) = &\left(\frac{M_p (2z)^p/(2+z)^p}{(2z)^p \psi_p(z)}\right)\left(\frac{e^{-(m-1/2)z} K_{p+m}([u-(m-\frac{1}{2})]z)}{K_{p+m}(uz)}\right) \\ &\cdot \left(\frac{u}{u-(m-\frac{1}{2})}\right)^{m+p}. \end{aligned} \tag{5.30}$$

By (1.6) the middle factor in (5.30) is finite for all z in $[0, \infty)$ and all $p \geqq 1$ and the same is also true for the last factor. By the previous argument, the ratio of the maxima of the absolute values of the numerator and denominator in the first factor in (5.30) over $[0, \infty)$ is also finite for each $p \geqq 1$. Furthermore, from (5.30) and (1.6) it follows that the error bound in (1.4) mimics the behavior of the first neglected term near both $z = 0$ and $z = \infty$.

Finally, we remark that the presence of factors like $[u-(m-\frac{1}{2})]^{-(m+p)}$ in the error bound (1.4) frequently occurs in asymptotics (see [3, p. 89, (9.02)], for example), and serves here as a reminder that two competing exponential factors appear in the integrand

of the integral representation (2.1), one growing like $e^{(m-1/2)t}$ for $m > \frac{1}{2}$, the other decaying like e^{-ut}.

REFERENCES

[1] A. ERDÉLYI, ED., *Higher Transcendental Functions*, Vol. 1, Bateman Manuscript Project, McGraw-Hill, New York, 1953.

[2] S. KARLIN, *Total Positivity*, Vol. 1, Stanford University Press, Stanford, CA, 1968.

[3] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[4] M. POURAHMADI, *Taylor expansion of* $\exp(\sum_{k=0}^{\infty} a_k z^k)$ *and some applications*, Amer. Math. Monthly, 91 (1984), pp. 303–308.

[5] J. RIORDAN, *An Introduction to Combinatorial Analysis*, Princeton University Press, Princeton, NJ, 1980.

[6] P. N. SHIVAKUMAR AND R. WONG, *Error bounds for a uniform asymptotic expansion of the Legendre function* $P_n^{-m}(\cosh z)$, Quart. Appl. Math., 46 (1988), pp. 474–488.

[7] F. URSELL, *Integrals with a large parameter: Legendre functions of large degree and fixed order*, Math. Proc. Cambridge Philos. Soc., 95 (1984), pp. 367–380.

[8] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, Cambridge, 1944.

SIAM J. MATH. ANAL.
Vol. 21, No. 2, pp. 536-549, March 1990
© 1990 Society for Industrial and Applied Mathematics
014

FUNCTIONAL INEQUALITIES FOR COMPLETE ELLIPTIC INTEGRALS AND THEIR RATIOS*

G. D. ANDERSON,† M. K. VAMANAMURTHY,‡ AND M. VUORINEN§

Abstract. Some functional inequalities satisfied by complete elliptic integrals of the first kind are obtained. These inequalities are sharp and generalize the functional identity of Landen. A related inequality is given for certain quotients of such integrals.

Key words. complete elliptic integral, quasiconformal mapping, functional inequality, Teichmüller ring, Grötzsch ring

AMOS(MOS) subject classifications. primary 33A25, 33A70; secondary 30C60

1. Introduction.

For $0<r<1$ the functions

$$K(r)=\int_0^{\pi/2}(1-r^2\sin^2 t)^{-1/2}\,dt,\quad K'(r)=K(r'),\quad r'=\sqrt{1-r^2}, \tag{1.1}$$

$$E(r)=\int_0^{\pi/2}(1-r^2\sin^2 t)^{1/2}\,dt,\quad E'(r)=E(r'),\quad r'=\sqrt{1-r^2} \tag{1.2}$$

are known as complete elliptic integrals of the first and second kind, respectively [BF], [Bo], [BB2], and their values are listed in standard tables (e.g., [AS], [Fr]). The special combinations

$$\mu(r)=\frac{\pi}{2}\frac{K'(r)}{K(r)},\quad \gamma(s)=\frac{2\pi}{\mu(r)},\quad r=\frac{1}{s} \tag{1.3}$$

are particularly important in quasiconformal analysis (cf. [LV]).

The elliptic integral $K(r)$ satisfies the following basic identities due to Landen [BF, #163.01, 164.02] (cf. [WW, p. 507]):

$$K\left(\frac{2\sqrt{r}}{1+r}\right)=(1+r)K(r),\qquad K\left(\frac{1-r}{1+r}\right)=\frac{1}{2}(1+r)K'(r), \tag{1.4}$$

while the function $\mu(r)$ satisfies the identities

$$\mu(r)\mu(\sqrt{1-r^2})=\frac{\pi^2}{4},\quad \mu(r)\mu\left(\frac{1-r}{1+r}\right)=\frac{\pi^2}{2},\quad \mu(r)=2\mu\left(\frac{2\sqrt{r}}{1+r}\right) \tag{1.5}$$

(cf. [LV, Forms. (2.7), (2.9), (2.3), pp. 60, 61]). The first identity in (1.5) follows directly from definition (1.1), while the other two follow from (1.4). It is also well known that

$$\log\frac{1}{r}<\mu(r)<\log\frac{4}{r} \tag{1.6}$$

for $0<r<1$ [LV, Form. (2.10), p. 61].

* Received by the editors January 25, 1988; accepted for publication (in revised form) March 7, 1989.

† Michigan State University, East Lansing, Michigan 48824. This author's work was supported in part by a grant from the National Science Foundation and in part by a grant from the Academy of Finland.

‡ University of Auckland, Auckland, New Zealand. This author's work was supported in part by a grant from the Academy of Finland.

§ University of Helsinki, Helsinki, Finland. This author's work was supported in part by the Alexander von Humboldt Foundation.

In this paper we study some properties of the functions $K(r)$ and $\mu(r)$. These special functions are important in the theory of quasiconformal mappings in the plane [LV] and in n-space [AVV1]-[AVV4], [Vu2]. In the present paper we prove certain functional inequalities for $K(r)$ and $\mu(r)$ by analytic arguments (mainly well-known properties of $K(r)$). Our work is motivated in part by certain functional inequalities, valid for higher-dimensional analogues of $\mu(r)$, which were proved in [Vu1] by geometric arguments. Our main results are as follows.

THEOREM 1.7. *For $r, a, b \in (0,1)$, the functions $K(r)$, $K'(r)$ satisfy the functional inequalities*

$$K(\sqrt{r}) < (1+r)K(r), \tag{1.8}$$

$$(1+r)K'(r) < 2K'(\sqrt{r}), \tag{1.9}$$

$$K\left(\frac{2\sqrt[4]{ab}}{\sqrt{(1+a)(1+b)}}\right) \leqq (1+\sqrt{ab})K(\sqrt{ab}), \tag{1.10}$$

$$2K'\left(\frac{2\sqrt[4]{ab}}{\sqrt{(1+a)(1+b)}}\right) \geqq (1+\sqrt{ab})K'(\sqrt{ab}). \tag{1.11}$$

There is equality in (1.10) *and* (1.11) *if and only if $a=b$, while* (1.8) *and* (1.9) *hold with equality at $r=0,1$, respectively, and* (1.9) *is asymptotically sharp at $r=0$.*

THEOREM 1.12. *For $a, b \in (0,1)$,*

$$\mu(a)+\mu(b) \leqq \mu\left(\frac{ab}{(1+\sqrt{1-a^2})(1+\sqrt{1-b^2})}\right) \leqq 2\mu(\sqrt{ab}).$$

In each case equality holds if and only if $a=b$.

Throughout this paper, for $t \in [0,1]$, t' will denote $\sqrt{1-t^2}$, as in (1.1) and (1.2). When the argument of the function is clear, we will frequently write K, K' and E, E' instead of $K(r)$, $K'(r)$ and $E(r)$, $E'(r)$. We will follow the relatively standard notation of [LV].

2. Properties of the function $K(r)$. The following differentiation formulas will be useful in our work.

LEMMA 2.1. *For $0<r<1$,*

$$(1) \qquad \frac{d}{dr}\frac{K'(r)}{K(r)} = \frac{-\pi}{2rr'^2K(r)^2},$$

$$(2) \qquad \mu'(r) = \frac{-\pi^2}{4rr'^2K(r)^2},$$

where $K(r)$, $\mu(r)$ are as in (1.1) *and* (1.3), *respectively.*

Proof. Formula (1) appears in [C, p. 217] and [E, p. 445]. It may also be obtained from the differentiation formula for $K(r)$ ([BF, #710.00], [Bo, p. 21]) and Legendre's relation ([BF, #110.10], [Bo, p. 25]). Formula (2) follows immediately from (1) and (1.3). □

THEOREM 2.2. *As a function of r,*

(1) *$K(r)$ is strictly increasing from $(0,1)$ onto $(\pi/2, \infty)$. Moreover, $f_1(r) \equiv K(r)+c\log r$ is increasing if and only if $c \geqq 0$; there is no c for which $f_1(r)$ is decreasing.*

(2) *$K(r)+\log r'$ is strictly decreasing from $(0,1)$ onto $(\log 4, \pi/2)$, and $K(r)+GG(\pi/4)\log r'$ is strictly increasing from $(0,1)$ onto $(\pi/2, \infty)$. Moreover, $f_2(r) \equiv K(r)+c\log r'$ is increasing if and only if $c \leqq \pi/4$ and decreasing if and only if $c \geqq 1$.*

(3) *$r'K(r)^2$ is strictly decreasing from $[0,1)$ onto $(0, \pi^2/4]$.*

(4) *Let* $f(r) = r' \exp(r^2K/(E - r'^2K))$, $0 < r < 1$, *and* $f(0) = e^2$, $f(1) = 4$. *Then* f *is strictly decreasing from* $[0, 1]$ *onto* $[4, e^2]$. *In particular, for* $0 < r < 1$,

$$4 < r' \exp(r^2K/(E - r'^2K)) < e^2,$$

and these bounds are sharp as r *tends to one, zero, respectively.*

(5) $K(r)/\log(4/r')$ *is strictly decreasing from* $(0, 1)$ *onto* $(1, \pi/\log 16)$, *and* $K(r)/\log(e^2/r')$ *is strictly increasing from* $(0, 1)$ *onto* $(\pi/4, 1)$. *Moreover,* $f_3(r) \equiv K(r)/\log(c/r')$ *is strictly decreasing if and only if* $0 < c \leqq 4$ *and strictly increasing if and only if* $c \geqq e^2$. *In particular, for* $0 < r < 1$,

$$\log(4/r') \leqq K(r) \leqq (\pi/\log 16)\log(4/r'),$$

$$(\pi/4)\log(e^2/r') \leqq K(r) \leqq \log(e^2/r').$$

(6) $r \exp(\pi K'/2K)$ *is strictly decreasing from* $(0, 1)$ *onto* $(1, 4)$.

(7) $r^{-2}(E - r'^2K)$ *is strictly increasing from* $(0, 1)$ *onto* $(\pi/4, 1)$.

(8) $K(r)K'(r)$ *is strictly decreasing on* $(0, 1/\sqrt{2}]$ *and strictly increasing on* $[1/\sqrt{2}, 1)$.

(9) $K(r)^{-2} + K'(r)^{-2}$ *has its absolute maximum on* $(0, 1)$ *at* $1/\sqrt{2}$.

Proof. The first statement in part (1) follows directly from definition (1.1). For the stronger statement we note by [BF, #710.00] that $f_1(r)$ is increasing on $(0, 1)$

$$\Leftrightarrow (E - r'^2K)/(rr'^2) + c/r \geqq 0, \qquad 0 < r < 1$$

$$\Leftrightarrow -c \leqq \inf\{(E - r'^2K)/r'^2 : 0 < r < 1\} = 0$$

$$\Leftrightarrow c \geqq 0.$$

Similarly, $f_1(r)$ cannot be decreasing since this would require $c \leqq -\sup\{(E - r'^2K)/r'^2 : 0 < r < 1\} = -\infty$.

The first monotonicity property in (2) was proved in [AV, Form. (2)], and from this the limits $f_2(0+)$ and $f_2(1-)$ when $c = \pi/4$ are clear. For the rest of (2) we note that $f_2'(r) \geqq 0$ on $(0, 1)$

$$\Leftrightarrow (E - r'^2K)/(rr'^2) - cr/r'^2 \geqq 0, \qquad 0 < r < 1$$

$$\Leftrightarrow c \leqq \inf\{(E - r'^2K)/r^2 : 0 < r < 1\} = \pi/4$$

and $f_2'(r) \leqq 0$

$$\Leftrightarrow c \geqq \sup\{(E - r'^2K)/r^2 : 0 < r < 1\} = 1 \text{ [AV, Form. (1)]}.$$

For (3), let $f(r) = r'K(r)^2$. By [BF, #710.00] we may write $f'(r) = K(r)g(r)/(rr')$, where $g(r) \equiv (E - r'^2K) + (E - K)$. But it is easy to see that $g(0) = 0$ and $g'(r) = rK - rE/r'^2 < 0$ on $(0, 1)$ [BF, #710.04, 710.05], hence $f'(r) < 0$ on $(0, 1)$. Clearly $f(0) = \pi^2/4$ and $f(1-) = 0$.

In part (4), if we let $g(r) = \log f(r)$ we get, after simplification [BF], $g'(r) = rK(E - r'^2K)^{-2}[(E - r'^2K) - (K - E)]$. Since $(E - r'^2K) = 0 = (K - E)$ when $r = 0$ and since $(d/dr)[(E - r'^2K) - (K - E)] = r'^{-2}(r'^2K - E) < 0$, it follows that $g'(r) < 0$ on $(0, 1)$ and that f is decreasing. The limit $f(0+) = e^2$ follows from [AV, Form. (1)]. For $f(1-) = 4$ it is sufficient to prove that $\lim_{r\to 1}[r^2/(E - r'^2K) - 1]K = 0$, since by [BF, #112.01] we know that $\lim_{r\to 1} r'\, e^K = 4$. First, as r tends to 1,

$$h(r) \equiv \left[\frac{r^2}{E - r'^2K} - 1\right]K = \frac{r^2}{E - r'^2K}\left[1 - \frac{E - r'^2K}{r^2}\right]K \sim [r^2 - E + r'^2K]K$$

by [AV, p. 62, line 1]. But $r'^2K^2 \to 0$ as $r \to 1$ by (3). So $h(r) \sim (r^2 - E)K$ as $r \to 1$. Then, by l'Hôpital's rule, [BF, #710], [AV, Form. (1)], and (3) we have

$$(r^2 - E)K = \frac{r^2 - E}{1/K} \sim \frac{2r - (E-K)/r}{-K^{-2}(E - r'^2K)/(rr'^2)} = \frac{[2 - (E-K)/r^2]r'^2K^2}{-(E - r'^2K)/r^2}$$

$$\sim [(E-K)/r^2 - 2](r'K)^2 \sim -r'^2K^3 = -\frac{(r'K^2)^2}{K} \to 0.$$

Parts (6) and (7) and the first monotonicity property in (5) were proved in (1), (3), and (4), respectively, of [AV]. The assertions about $K(r)/\log(e^2/r')$ in (5) follow from the properties of $K(r)/\log(4/r')$. The remainder of (5) follows from (4) since $f_3'(r) \leqq 0$ on (0, 1)

$$\Leftrightarrow (\log(c/r'))^2 f_3'(r) = (\log(c/r'))(E - r'^2K)/(rr'^2) - rK/r'^2 \leqq 0, \qquad 0 < r < 1$$

$$\Leftrightarrow c \leqq \inf\{r' \exp((r^2K)/(E - r'^2K)) : 0 < r < 1\} = 4,$$

and

$$f_3'(r) \geqq 0, \qquad 0 < r < 1 \Leftrightarrow c \geqq \sup\{r' \exp(r^2K/(E - r'^2K)) : 0 < r < 1\} = e^2.$$

For (8), by [BF, #710.00] we have

$$\frac{d}{dr}(K(r)K'(r)) = \frac{1}{rr'^2}[F(r) - F(r')],$$

where

$$F(r) = K'(E - r'^2K) = r^2K' \int_0^{\pi/2} \frac{\cos^2 t}{\sqrt{1 - r^2 \sin^2 t}}\, dt.$$

It follows from (3) that $r'^2K(r)$ is strictly decreasing on [0, 1), and hence that $r^2K'(r)$ is strictly increasing there. Thus $F(r)$, as the product of two positive strictly increasing functions, is strictly increasing on (0, 1). Hence $F(r) - F(r')$ is negative when $0 < r < 1/\sqrt{2}$ and positive when $1/\sqrt{2} < r < 1$.

Finally, for (9), let $G(r) = K(r)^{-2} + K'(r)^{-2}$, $0 < r < 1$. Then by [BF, #710.00] we have

$$G'(r) = \frac{-2}{rr'^2}\left[\frac{E - r'^2K}{K^3} - \frac{E' - r^2K'}{K'^3}\right],$$

which is zero when $r = 1/\sqrt{2}$. It is easy to see that $G'(r) = 0$ if and only if

$$\frac{r^2K'^3}{r'^2K^3} = \frac{\int_0^{\pi/2} ((\cos^2 t)/\sqrt{1 - r'^2 \sin^2 t})\, dt}{\int_0^{\pi/2} ((\cos^2 t)/\sqrt{1 - r^2 \sin^2 t})\, dt}.$$

But the right side of this last equation is obviously decreasing for $r \in (0, 1)$, while by (3) the left side is increasing there. Since $G(0+) = G(1-) = 4/\pi^2 \approx 0.4053 < 0.5818 \approx 2/K(1/\sqrt{2})^2 = G(1/\sqrt{2})$, we have $\max_{0<r<1} G(r) = G(1/\sqrt{2})$. □

COROLLARY 2.3. *For* $0 < r < 1$, $K(r) \leqq (\pi/2)\log(e/r')$.

Proof. From 2.2(5) we have

$$K(r) - \frac{\pi}{2}\log\left(\frac{e}{r'}\right) \leqq \frac{\pi}{\log 16}\log\frac{4}{r'} - \frac{\pi}{2}\log\frac{e}{r'} = \frac{\pi \log(1/r')}{\log 16}(1 - \log 4) < 0. \qquad \square$$

Remarks 2.4. (1) Theorem 2.2(3) is a best possible result, in the sense that if $c > 0$ then $f(r) \equiv r'K(r)^{2+c}$ is not monotone on [0, 1). For we can then show that $f'(r) = K^{1+c}g(r)/(rr')$, where $g(r) \equiv -r^2K + (2+c)(E - r'^2K)$. Moreover, $g'(r) = rh(r)/r'^2$, where $h(r) \equiv (c+1)r'^2K - E$. Then $h(0) = c\pi/2 > 0$ and $h(1) = -1 < 0$. Thus there exists $\delta > 0$ such that $f'(r) > 0$ on $(0, \delta)$ and $f'(r) < 0$ on $(1 - \delta, 1)$.

(2) By combining Theorem 2.2(5) with the first transformation in (1.4) we may easily show that K satisfies the inequality

$$K(r) \geqq \frac{1}{1+r} \log\left(4\frac{1+r}{1-r}\right), \qquad 0<r<1. \tag{2.5}$$

This procedure may be repeated; the second time, for example, it leads to the inequality

$$K(r) \geqq \frac{2}{(1+\sqrt{r})^2} \log\left(2\frac{1+\sqrt{r}}{1-\sqrt{r}}\right), \qquad 0<r<1. \tag{2.6}$$

2.7. *Proof of Theorem* 1.7. First we note that $\sqrt{r}<2\sqrt{r}/(1+r)$ for $r\in(0,1)$. Then since K is strictly increasing on $[0,1)$, (1.8) follows from the first identity in (1.4). Since K' is strictly decreasing on $(0,1]$, the second identity in (1.4) implies (1.9). The equality statements for (1.8) and (1.9) follow from the fact that $K(0)=\pi/2=K'(1)$, while the assertion about asymptotic sharpness is a consequence of [BF, #112.01].

For (1.10) we first observe that $(1+a)(1+b)\geqq(1+\sqrt{ab})^2$ for $a, b>0$. Then it follows from the first identity in (1.4) that

$$K\left(\sqrt{\frac{4\sqrt{ab}}{(1+a)(1+b)}}\right) \leqq K\left(\frac{2\sqrt[4]{ab}}{1+\sqrt{ab}}\right) = (1+\sqrt{ab})K(\sqrt{ab}). \tag{2.8}$$

Again, since K' is decreasing and $(1+a)(1+b)\geqq(1+\sqrt{ab})^2$ for $a, b>0$, we have by (1.4)

$$K'\left(\frac{2\sqrt[4]{ab}}{\sqrt{(1+a)(1+b)}}\right) \geqq K'\left(\frac{2\sqrt[4]{ab}}{1+\sqrt{ab}}\right) = K\left(\frac{1-\sqrt{ab}}{1+\sqrt{ab}}\right) = \frac{1+\sqrt{ab}}{2} K'(\sqrt{ab}).$$

Since K is strictly increasing, there is equality in (1.10) if and only if the first two arguments in (2.8) are equal, and it is easy to show that this occurs if and only if $a=b$. Similarly, there is equality in (1.11) if and only if $a=b$. □

COROLLARY 2.9. *For* $a, b\in(0,1)$,

$$\frac{(1+a')(1+b')}{2(1+a'b')} < \frac{K(ab/((1+a')(1+b')))}{K(ab/(1+a'b'))} < 1.$$

Proof. The upper bound is trivial since $(1+a')(1+b')>1+a'b'$. For the lower bound, we have

$$K\left(\frac{ab}{(1+a')(1+b')}\right) = K\left(\frac{\sqrt{(1-a')(1-b')}}{\sqrt{(1+a')(1+b')}}\right) > K\left(\frac{(1-a')(1-b')}{(1+a')(1+b')}\right),$$

and the second identity in (1.4) yields

$$\begin{aligned} K\left(\frac{(1-a')(1-b')}{(1+a')(1+b')}\right) &= K\left(\frac{1-(a'+b')/(1+a'b')}{1+(a'+b')/(1+a'b')}\right) \\ &= \frac{1}{2}\left(1+\frac{a'+b'}{1+a'b'}\right) K'\left(\frac{a'+b'}{1+a'b'}\right) \\ &= \frac{1}{2}\frac{(1+a')(1+b')}{1+a'b'} K\left(\frac{ab}{1+a'b'}\right). \end{aligned}$$

□

3. Properties of the function μ. In [AVV3] we obtained some properties of the n-dimensional analogue of μ for all $n\geqq 2$. Now, using the explicit formula for $\mu(r)$ in (1.3), we obtain some stronger properties for the case $n=2$.

3.1. *Proof of Theorem* 1.12. The second inequality is a special case of [AVV3, Form. (1.24)]. Next, by [AVV3, Form. (1.22)] and the third identity in (1.5) we have

$$\mu\left(\frac{2\sqrt{r}}{1+r}\right)+\mu\left(\frac{2\sqrt{s}}{1+s}\right)\leqq\mu(\sqrt{rs}).$$

Setting $a=2\sqrt{r}/(1+r)$ and $b=2\sqrt{s}/(1+s)$ then gives the first inequality.

Since this argument shows that the theorem is equivalent to [AVV3, Form. (1.22)] when $n=2$, it is sufficient to prove the equality statement for [AVV3, Form. (1.22)] with $n=2$. But by the third identity in (1.5) we have, with $a=\cos(2x)$, $b=\cos(2y)$, $x, y\in[0, \pi/4]$,

$$\mu(a)+\mu(b)=2\mu(\sqrt{ab})\Leftrightarrow\sqrt{(1+a')(1+b')}+\sqrt{(1-a')(1-b')}=2$$
$$\Leftrightarrow\cos(x-y)=1\Leftrightarrow x=y\Leftrightarrow a=b. \qquad \square$$

Remark 3.2. Alternatively, we could prove Theorem 1.12 by using derivatives of elliptic integrals.

THEOREM 3.3. *For* $0<r<1$,

$$\mu(r^2)\mu\left(\left(\frac{1-r}{1+r}\right)^2\right)=\pi^2 \tag{3.4}$$

and for each integer $k\geqq 2$,

$$2^{k-1}\pi^2<\mu(r^{2^k})\mu\left(\left(\frac{1-r}{1+r}\right)^{2^k}\right)<2^{2k-1}\pi^2. \tag{3.5}$$

The first bound in (3.5) *is asymptotically sharp as r tends either to zero or to one.*

Proof. First, by the third and second identities in (1.5),

$$\mu(r^2)\mu\left(\left(\frac{1-r}{1+r}\right)^2\right)=2\mu(r^2)\mu\left(\frac{1-r^2}{1+r^2}\right)=\pi^2.$$

Next, suppose $k=2$. Then by (1.5), (3.4), and the fact that μ is strictly decreasing we have

$$\mu(r^4)\mu\left(\left(\frac{1-r}{1+r}\right)^4\right)=2\mu(r^4)\mu\left(\frac{(1-r^2)^2}{1+6r^2+r^4}\right)>2\mu(r^4)\mu\left(\left(\frac{1-r^2}{1+r^2}\right)^2\right)=2\pi^2.$$

Next, for the upper bound, from Theorem 1.12 and (3.4) we get

$$\mu(r^4)\mu\left(\left(\frac{1-r}{1+r}\right)^4\right)\leqq 4\mu(r^2)\mu\left(\left(\frac{1-r}{1+r}\right)^2\right)=4\pi^2<8\pi^2.$$

Now suppose that $n=2^k$, where $k\geqq 2$ is an integer for which (3.5) is true. Then by the third identity in (1.5)

$$\mu\left(\left(\frac{1-r}{1+r}\right)^{2n}\right)=2\mu\left(\frac{2(1-r^2)^n}{(1+r)^{2n}+(1-r)^{2n}}\right)>2\mu\left(\left(\frac{1-r^2}{1+r^2}\right)^n\right),$$

since

$$\binom{2n}{2l}\geqq\binom{n}{l}$$

for $0\leqq l\leqq n$ with equality only if $l=0$ or n. The first inequality in (3.5) now follows by the second identity in (1.5) and by mathematical induction.

Finally, for the upper bound in (3.5) we again use induction:

$$\mu(r^{2n})\mu\left(\left(\frac{1-r}{1+r}\right)^{2n}\right)=\mu((r^n)^2)\mu\left(\left(\left(\frac{1-r}{1+r}\right)^n\right)^2\right)\leqq 2\mu(r^n)\cdot 2\mu\left(\left(\frac{1-r}{1+r}\right)^n\right)$$
$$\leqq 4\cdot 2^{2k-1}\pi^2$$
$$=2^{2(k+1)-1}\pi^2.$$

The asymptotic limits are a consequence of [AVV3, Lemma 2.6(2), (4)]. □

THEOREM 3.6. *If $0<r<1$ and $m=2^k$, where $k\geqq 2$ is an integer, then*

$$m<\frac{\mu(((1-r)/(1+r))^m)}{\mu((1-r^m)/(1+r^m))}<m^2.$$

Proof. The proof follows by (3.5) and the second identity in (1.5). □

THEOREM 3.7. *For $0<r<1$, $r'\log(4/r)<\mu(r)<\log(4/r)$. These estimates are sharp as r tends to zero.*

Proof. The second estimate is contained in (1.6). The first follows from Theorem 2.2(3),(5), which says that

$$\frac{\mu(r)}{r'\log(4/r)}=\frac{\pi}{2r'K(r)}\frac{K'(r)}{\log(4/r)}$$

is strictly increasing from $(0,1)$ onto $(1,\infty)$. □

COROLLARY 3.8. *For $a, b\in(0,1)$,*

$$\mu(a)<2\mu(\sqrt{a}), \tag{3.9}$$

$$\mu(\sqrt{ab})\leqq 2\mu\left(\frac{2\sqrt[4]{ab}}{\sqrt{(1+a)(1+b)}}\right). \tag{3.10}$$

There is equality in (3.10) if and only if $a=b$.

Proof. For (3.9), divide (1.9) by (1.8). For (3.10), divide (1.11) by (1.10). The equality in (3.10) follows from the third identity in (1.5). □

4. Properties of the graph of μ. The following lemma will be needed in this section.

LEMMA 4.1. *Let $g:[0,1]\to[0,\infty)$ be continuous and (strictly) convex with $g(0)=0$, and let g be differentiable on $(0,1)$. Then $g(r)/r$ is (strictly) increasing on $(0,1]$.*

Proof. By the Mean Value Theorem,

$$\frac{d}{dr}\left(\frac{g(r)}{r}\right)=\frac{1}{r}\left(g'(r)-\frac{g(r)}{r}\right)=\frac{1}{r}(g'(r)-g'(r_1))\geqq 0,$$

where $0<r_1<r$. □

The function introduced in the next lemma behaves very much like μ (cf. [AVV3, Lemma 2.6, Cor. 2.8] as well as Theorem 4.3 below) and will be useful in the study of μ.

LEMMA 4.2. *For $0<r<1$ let $m(r)\equiv(2/\pi)r'^2K(r)K'(r)$. Then:*

(1) *$m(r)+\log r$ is a strictly decreasing concave function from $(0,1)$ onto $(0,\log 4)$.*
(2) *$m(r)/\log(1/r)$ is strictly increasing from $(0,1)$ onto $(1,\infty)$.*
(3) *$m(r)/\log(4/r)$ is strictly decreasing from $(0,1)$ onto $(0,1)$.*
(4) *$(\log(4/r)-m(r))/r$ is strictly increasing from $(0,1)$ onto $(0,\infty)$.*
(5) *$\log(1/r)<m(r)<r'\log(4/r)<\mu(r)<\log(4/r)$.*

Proof. For (1) let $f(r)=m(r)+\log r$. Then by differentiation [BF, #710.00] and by Legendre's relation [BF, #110.10] we have

$$-f'(r)=\frac{4}{\pi r}K'(K-E)=\frac{4}{\pi}rK'\int_0^{\pi/2}\frac{\sin^2 t}{\sqrt{1-r^2\sin^2 t}}\,dt,$$

which is positive and increasing on $(0, 1)$, while $f(1-)=0$ by Theorem 2.2 and [BF, #111.02]. Moreover, by [BF, #112.01] or [Bo, p. 21, Form. (20)] and the fact that

$$\frac{2}{\pi} r'^2 K - 1 = -r^2 + O(r^2)$$

([BF, #900.00], [Bo, p. 21]), we have

$$f(0+) = \frac{2}{\pi} \lim_{r\to 0} r'^2 K(K' + \log r) + \lim_{r\to 0} \left(\log \frac{1}{r}\right)\left(\frac{2}{\pi} r'^2 K - 1\right) = \log 4.$$

For part (2), setting $g(r) = m(r)/\log(1/r)$, differentiating, and using Legendre's relation [BF, #110.10] we get

$$\frac{\pi}{2} r \left(\log \frac{1}{r}\right)^2 g'(r) = \left(2E'K - \frac{\pi}{2}\right) F(r),$$

where

$$F(r) \equiv \log r + \frac{r'^2 KK'}{2E'K - \pi/2}.$$

Similarly, by [BF, #710.00, 710.02, 110.10],

$$F'(r) = \frac{2KK'}{r} \frac{(K-E)E' - r^2 KK'}{(2E'K - \pi/2)^2},$$

where

$$(K-E)E' - r^2 KK' \leqq r^2 K(E' - K') < 0$$

since $K - E < r^2 K$ and $E' < K'$ on $(0, 1)$ [BF, #110.06, 110.07]. Thus $F'(r) < 0$, so that $F(r) > F(1) = 0$ for $0 < r < 1$. Since $2E'K > \pi$, this implies that $g'(r) > 0$ on $(0, 1)$. The limit as r tends to zero follows from [BF, #111.02, 112.01]; the limit as r tends to one follows from l'Hôpital's rule.

From (1) it follows that $\log(4/r) - m(r)$ is a strictly increasing and convex function from $(0, 1)$ onto $(0, \log 4)$. Then (3) follows when we divide by $\log(4/r)$, and (4) follows from Lemma 4.1.

Finally, because of (2), (1.6), and Theorem 3.7, we need only prove the second inequality in (5). For this we define $f(r) = \log(4/r) - (2/\pi) r' KK'$. Then writing

$$f(r) = -(K' - \log(4/r)) + K'(1-r') + r'K'(1 - (2/\pi)K)$$

and using Theorem 2.2(2) and [BF, #900.00] gives $f(0+) = 0$, while

$$f'(r) = -\frac{1}{r} + \frac{1}{rr'} + \frac{2K'}{\pi rr'}[(K-E) - (E - r'^2 K)]$$

by [BF, #710.00] and Legendre's relation. With $g(r) \equiv (K-E) - (E - r'^2 K)$, we have $g(0) = 0$ and

$$g'(r) = \frac{r}{r'^2} - rK = \frac{r}{r'^2}(E - r'^2 K) > 0.$$

Thus $g(r) > 0$ on $(0, 1)$, hence $f'(r) > 0$ on $(0, 1)$, and (5) follows. □

THEOREM 4.3. (1) *The function $\mu(r)$ is strictly decreasing, has exactly one inflection point on $(0, 1)$ and satisfies $\mu'(0+) = -\infty = \mu'(1-)$.*

(2) *The function $g(r) \equiv \mu(r) + \log r$ is strictly decreasing and concave on $(0, 1)$ and satisfies $g'(0+) = 0$, $g'(1-) = -\infty$.*

(3) *The function $h(r) \equiv \mu(r) + \log(r/r')$ is strictly increasing and convex on $(0, 1)$ and satisfies $h'(0+) = 0$, $h'(1-) = \infty$.*

(4) *The function* $\mu(r)/\log(1/r)$ *is strictly increasing but is neither convex nor concave on* $(0,1)$.

(5) *The function* $\mu(r)/\log(4/r)$ *is strictly decreasing and concave on* $(0,1)$.

(6) *The function* $1/\mu(r)$ *is strictly increasing and has exactly one inflection point on* $(0,1)$.

Proof. The monotonicity in parts (1), (2), and (6) is well known ([LV, Lemma 6.3], [G1, Lemma 6], [A, §§ 8, 9]; cf. [AVV3, Proof of Lemma 2.6(1)]), and in the others it follows from [AVV3, Lemma 2.6]. The limits $\mu'(0+) = -\infty = \mu'(1-)$ follow immediately from Lemma 2.1 and Theorem 2.2. Next, let $f(r) = r(1-r^2)K(r)^2$ be the denominator function in the formula for $\mu'(r)$ in Lemma 2.1. By [BF, #710.00],

$$f'(r) = K(r)[-(1+r^2)K(r)+2E(r)],$$

where E is as in (1.2). Then by [BF, #111.02, 111.03, 111.05] or [Bo, p. 20] we have $f'(0+) = \pi^2/4$ and $f'(1-) = -\infty$. The intermediate value theorem implies that there exists at least one $r_0 \in (0,1)$ such that $f'(r_0) = 0 = \mu''(r_0)$, and there is at most one such r_0 because $f'(r)/K(r)$ is strictly decreasing. Thus μ is convex on $(0, r_0]$ and concave on $[r_0, 1)$.

For part (2), using Lemma 2.1 and [BF, #112.01] (cf. [Bo, p. 21]), we see that $g'(1-) = -\infty$, while Lemma 2.1, [BF, #710.00], l'Hôpital's rule, and [BF, #710.05] give

$$\lim_{r\to 0} g'(r) = \lim_{r\to 0} \frac{4}{\pi^2 r}(E(r)-K(r)) = \lim_{r\to 0} \frac{-4rE(r)}{\pi r'^2} = 0.$$

Then by Lemma 2.1, [BF, #710.00], and the fact that $K(r) \geqq \pi/2$,

$$g''(r) \leqq \frac{\pi^2}{4r^2r'^4K^3}[(-2+r^2-r^4)K+2E].$$

Let $F(r) = (-2+r^2-r^4)K+2E$. Then $F(0) = 0$ by [BF, #111.02], while

$$r'^2F'(r) = -r[E-r'^2K+r^2(E+3r'^2K)] < 0$$

by [AV, Form. (1)].

The limits in part (3) follow from part (2). Next, by Lemma 2.1,

$$h'(r) = \frac{1}{rr'^2}\left(\frac{-\pi^2}{4K(r)^2}+1\right) > 0, \qquad 0<r<1,$$

since $K(r) > \pi/2$. For convexity it is sufficient to show that $G(r) \equiv (1-\pi^2/(4K(r)^2))/r$ is increasing. It is easy to show that $G'(r) > 0$ if and only if $(2E(r)-r'^2K(r))-4K(r)^3r'^2/\pi^2 > 0$. But this follows from Theorem 2.2 and the fact that

$$r\frac{d}{dr}(2E(r)-r'^2K(r)) = E(r)-r'^2K(r) > 0$$

for $0<r<1$ [BF, #710.04, 710.02]. Hence part (3) follows.

Next, for part (4),

$$\frac{d}{dr}\left(\frac{\mu(r)}{\log(1/r)}\right) = \frac{\pi^2((2/\pi)r'^2KK'-\log(1/r))}{4rr'^2K^2(\log(1/r))^2}.$$

By Lemma 4.2(1) this has limit ∞ as r tends to zero, while Lemma 4.2(2) and Theorem 2.2 show that it tends to ∞ as r tends to one; thus part (4) follows.

For part (5) we write

$$-\frac{d}{dr}\left(\frac{\mu(r)}{\log(4/r)}\right) = \frac{\pi^2}{4}\frac{1}{r'^2K^2(\log(4/r))^2}\cdot\frac{\log(4/r)-(2/\pi)rr'^2KK'}{r},$$

which is increasing by Theorem 2.2 and Lemma 4.2(4).

Finally, for part (6), by Lemma 2.1 and [BF, #710.00] we have

$$\frac{d^2}{dr^2}\frac{1}{\mu(r)}=\frac{2\pi}{r^2r'^4K'(r)^3}[2E'(r)-r'^2K'(r)].$$

Since $2E'(r)-r'^2K'(r)$ is a strictly increasing function from $(0,1)$ onto $(-\infty,\pi)$, part (6) follows by the intermediate value theorem. □

Remarks 4.4. (1) The monotonicity of the function in Theorem 4.3(3) is stated by Gehring in [G2].

(2) Although $\mu(r)$ has been shown above to be neither convex nor concave on $(0,1)$, we now establish the interesting fact that the function $\mu(1/s)$ is concave and its reciprocal is convex on $(1,\infty)$.

THEOREM 4.5. *As functions of s on* $(1,\infty)$:

(1) $\mu(1/s)$ *is concave*;

(2) $g(s)\equiv\mu(1/s)/\log s$ *is convex*;

(3) $h(s)\equiv\mu(1/s)/\log(4s)$ *is concave.*

Proof. For part (1), let $f(s)=\mu(r)$, where $s=1/r$. Then by Lemma 2.1,

$$f'(s)=\frac{\pi^2 r}{4r'^2K(r)^2};$$

by Theorem 2.2 this is an increasing function of r, hence a decreasing function of s.

For part (2), set $r=1/s$ and $F(r)=g(s)$. Then

$$-g'(s)=-F'(r)\frac{dr}{ds}=\frac{\pi^2 r((2/\pi)r'^2KK'-\log(1/r))}{4r'^2K^2(\log(1/r))^2}$$

$$=\frac{\pi^2 r}{4r'^2K^2\log(1/r)}\left[\frac{(2/\pi)r'^2KK'}{\log(1/r)}-1\right],$$

which, by Lemma 4.2(2), is increasing as a function of r, and hence is decreasing as a function of s.

For part (3), set $r=1/s$ and $G(r)=h(s)$. Then

$$h'(s)=G'(r)\frac{dr}{ds}=\frac{\pi^2 r(\log(4/r)-(2/\pi)r'^2KK')}{4r'^2K^2(\log(4/r))^2},$$

which, by Theorem 2.2 and Lemma 4.2(1), is increasing as a function of r and hence decreasing as a function of s. □

LEMMA 4.6. *Let* $f\colon I\to(0,\infty)$ *be twice differentiable and concave on some interval* $I\subset R^1$. *Then* $1/f$ *is convex on* I.

Proof. Let $g(x)=1/f(x)$. Then

$$g''(x)=\frac{2(f'(x))^2}{f(x)^3}-\frac{f''(x)}{(f(x))^2}\geqq 0$$

since $f''(x)\leqq 0$. □

COROLLARY 4.7. *The function* $\gamma(s)$ *is decreasing and strictly convex on* $(1,\infty)$.

Proof. The proof is an immediate consequence of Theorem 4.5(1) and Lemma 4.6. □

Next we obtain estimates for $\mu(r)$ that improve on the inequalities

$$\log((1+\sqrt{r'})^2/r)<\mu(r)<\log(2(1+r')/r) \tag{4.8}$$

due to Lehto and Virtanen [LV, p. 62].

THEOREM 4.9. *For* $0<r<1$,

$$\frac{1}{2}\log\frac{1+\sqrt{r'}+2\sqrt[4]{r'}}{1-\sqrt{r'}}<\mu(r)<\frac{1}{2}\log\frac{1+\sqrt{r'}+\sqrt{2(1+r')}}{1-\sqrt{r'}}.$$

Proof. By means of a Möbius transformation we may map the plane Grötzsch ring $B^2\setminus[0, r]$ onto the symmetric ring $D = B^2(R)\setminus[-1, 1]$, where $R = (1+r')/r$, and by conformal invariance we have $\mu(r) = \operatorname{mod} D$. Next, the function $w = \frac{1}{2}(\zeta + 1/\zeta)$ maps the circular annulus $A = \{\zeta: 1 < |\zeta| < \rho\}$ conformally onto the elliptical ring $E(\rho)$ consisting of the set $\{w = u+iv: u^2/(\rho+1/\rho)^2 + v^2/(\rho-1/\rho)^2 < \frac{1}{4}\}$ minus a slit $[-1, 1]$ in the u-axis. If we take $\rho_1 + 1/\rho_1 = 2R = \rho_2 - 1/\rho_2$, we have $E(\rho_1) \subset D \subset E(\rho_2)$, and hence $\operatorname{mod} E(\rho_1) < \mu(r) < \operatorname{mod} E(\rho_2)$. Solving for ρ_1 and ρ_2 in terms of R and using conformal invariance we obtain

$$\log \frac{1+r'+\sqrt{2r'(1+r')}}{r} < \mu(r) < \log \frac{1+r'+\sqrt{2(1+r')}}{r}. \tag{4.10}$$

As in [LV] we may improve (4.10) by combining these inequalities with the third identity in (1.5), which is equivalent to $\mu(r) = \frac{1}{2}\mu((1-r')/(1+r'))$. In doing this we obtain the desired estimates. It follows from the elementary inequalities $(1+x)^2 < 2(1+x^2)$ and $(1+x^2)^3 < (1+x)^2(1+x^4)$ for all $x \in (0, 1)$ that the upper and lower estimates, respectively, in the theorem are better than those in (4.8). □

4.11. *Figures.* To illustrate some of the results in this paper, we computed $\mu(r)$ numerically for $0 < r < 1$, using the recursive method of Gauss [BB2], [To] (cf. [Tr], [As]). Figures 1-3 show graphs of several functions related to μ that illustrate Theorems 4.3 and 4.5.

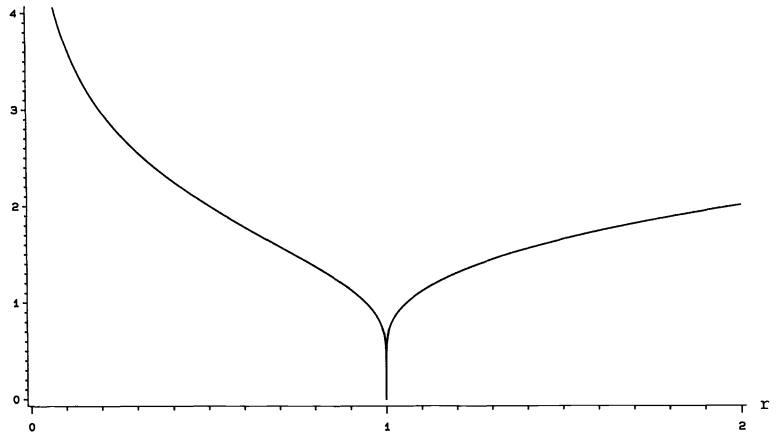


FIG. 1. $\mu(r)$, $0 < r \leqq 1$, *and* $\mu(1/r)$, $r > 1$.

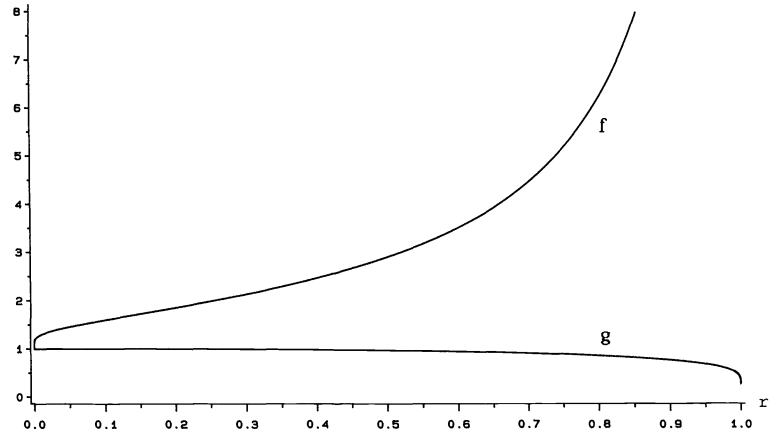


FIG. 2. $f(r) = \mu(r)/\log(1/r)$, $g(r) = \mu(r)/\log(4/r)$, $0 < r < 1$.

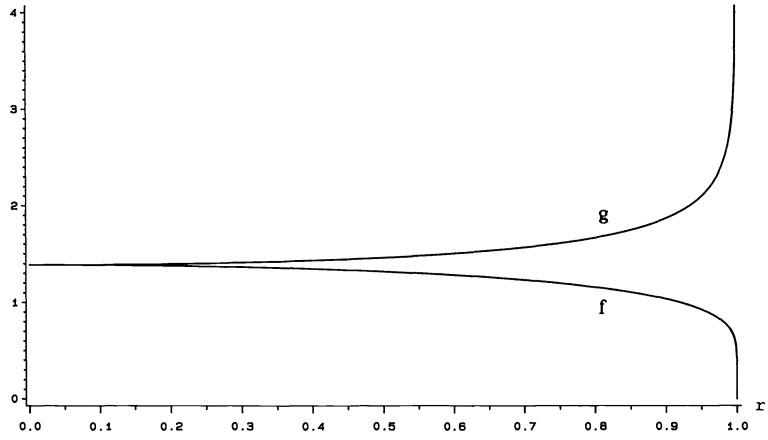


FIG. 3. $f(r)=\mu(r)+\log r$, $g(r)=\mu(r)+\log(r/r')$.

5. Properties of the Teichmüller capacity. The conformal capacity of the Teichmüller extremal ring is denoted here by $\tau(t)$, $t>0$. It is related to the capacity of the plane Grötzsch ring (see (1.3)) by the functional identity [G1, § 18]

$$\gamma(s)=2\tau(s^2-1). \tag{5.1}$$

In this section we study inequalities and limits for ratios of such Teichmüller capacities.

THEOREM 5.2. *For $1<A<\infty$ let $f(s)=\tau(s)/\tau(As)$ for all $s\in(0,\infty)$. Then f is strictly increasing on $(0,1/\sqrt{A}]$ and strictly decreasing on $[1/\sqrt{A},\infty)$ with $\lim_{s\to 0} f(s)=1=\lim_{s\to\infty} f(s)$. Consequently, $1\leqq f(s)\leqq f(1/\sqrt{A})$ for all $s\in(0,\infty)$.*

Proof. If we set $a=1/\sqrt{s+1}$, $b=1/\sqrt{As+1}$, then by (5.1) and the fact that μ is strictly decreasing we have $f(s)=\mu(b)/\mu(a)>1$.

Next, it follows easily from Lemma 2.1 that

$$\mu^2(a)f'(s)=\frac{\pi^3a^2}{16(1-a^2)K^2(a)K^2(b)}[K(a)K'(a)-K(b)K'(b)].$$

In the case where $0<s<1/A$, i.e., $1/\sqrt{2}<b<a$, it follows from Theorem 2.2(8) that $f'(s)>0$, so that f increases. If $1<s<\infty$, i.e., $0<b<a<1/\sqrt{2}$, then by the same theorem we have $f'(s)<0$, so that f is decreasing on $[1,\infty)$.

Finally, we consider the interval $[1/A,1]$. We note first that, by the definition of f and the first identity in (1.5),

$$f(1/A)=\frac{\mu(1/\sqrt{2})}{\mu(\sqrt{A/(A+1)})}=\frac{\mu(1/\sqrt{A+1})}{\mu(1/\sqrt{2})}=f(1).$$

Hence by the Mean Value Theorem there exists $s_0\in(1/A,1)$ such that $f'(s_0)=0$. Since $A>1$, $f'(s_0)=0$ if and only if $K(a)K'(a)=K(b)K'(b)$ if and only if $b=a'$ if and only if $s_0=1/\sqrt{A}$. Thus f has a unique critical point $s_0=1/\sqrt{A}\in(1/A,1)$, and hence $f(s_0)$ is either the maximum or minimum of $f(s)$ for $s\in[1/A,1]$.

We need only compare $f(s_0)$ with $f(1)=f(1/A)$. Setting $x=\sqrt{A}>1$, by (1.5) we have

$$f(s_0)=f\left(\frac{1}{x}\right)=\frac{\mu(1/\sqrt{x+1})}{\mu(\sqrt{x(x+1)})}=\frac{4}{\pi^2}(\mu(1/\sqrt{x+1}))^2=\frac{4}{\pi^2}(\mu(y))^2,$$

where $y = 1/\sqrt{x+1} \in (0, 1/\sqrt{2})$. Next,

$$f(1) = \frac{\mu(1/\sqrt{A+1})}{\mu(1/\sqrt{2})} = \frac{2}{\pi}\mu(1/\sqrt{x^2+1}) = \frac{2}{\pi}\mu(z),$$

where $z = 1/\sqrt{x^2+1}$. We let

$$g(x) = \frac{\mu^2(1/\sqrt{x+1})}{\mu(1/\sqrt{x^2+1})} = \frac{\mu^2(y)}{\mu(z)},$$

noting that $0 < z \leqq y \leqq 1/\sqrt{2}$ for $x \in [1, \infty)$. Clearly $g(1) = \pi/2$, and we need only prove that g is increasing on $[1, \infty)$. Differentiating and simplifying, we get

$$\mu^2(z)g'(z) = \frac{\pi^3}{8}\frac{K'(y)}{x^2K^3(y)K^2(z)}[K(z)K'(z) - K(y)K'(y)],$$

which is positive since $0 < z < y < 1/\sqrt{2}$ for all $x \in (1, \infty)$. Hence g is increasing on $[1, \infty)$, $g(x) > g(1) = \pi/2$ for all $x \in (1, \infty)$, and we conclude that $f(1/\sqrt{A}) > f(1) = f(1/A)$. If f' were negative at some point of $(1/A, 1/\sqrt{A})$, then by continuity f' would have another zero in this interval. This contradiction shows that f is strictly increasing on $[1/A, 1/\sqrt{A}]$. Similarly, f is strictly decreasing on $[1/\sqrt{A}, 1]$.

Finally, the limit as s tends to zero follows from (5.1) and [AVV3, Lemma 2.6(4)], and the limit as s tends to ∞ follows from (5.1) and [AVV3, Lemma 2.6(2)]. □

Remark 5.3. Theorem 5.2 enables us to replace the numerical bound 1.172 in [LeVu, § 1.2] by the exact constant $\tau(1/\sqrt{2})/\tau(\sqrt{2}) = (4/\pi^2)\mu^2(\sqrt{\sqrt{2}-1}) = 1.1712\cdots$.

6. Concluding remarks. Complete elliptic integrals have been tabulated and compared to other functions during the past century (cf. [LF, pp. 208–213]). The usual comparison function for $K(r)$ is $\log(4/\sqrt{1-r^2})$, and we relate our results to some earlier recent work on this topic.

In 1984 Borwein and Borwein [BB1, p. 356, Prop. 2] proved that

$$\left| K(r) - \log\frac{4}{\sqrt{1-r^2}} \right| \leqq 4(1-r^2)K(r). \tag{6.1}$$

In 1985 Carlson and Gustafson [CG, p. 1072, Form. (1.1)] went on to show that

$$1 < \frac{K(r)}{\log(4/\sqrt{1-r^2})} < \frac{4}{(3+r^2)} \tag{6.2}$$

for $r \in (0, 1)$. Note that the quantity between the absolute value bars in (6.1) is positive by Theorem 2.2(2) and that (6.2) yields the following improved upper bound for it:

$$0 < K(r) - \log\frac{4}{\sqrt{1-r^2}} \leqq \frac{1-r^2}{3+r^2}\log\frac{4}{\sqrt{1-r^2}} \leqq \frac{1-r^2}{3+r^2}K(r). \tag{6.3}$$

Remarks 6.4. (1) Theorem 2.2(5) strengthens the Carlson–Gustafson inequality (6.2) for $0 \leqq r < \sqrt{(4/\pi)\log 16 - 3} = 0.728\cdots$.

(2) Some computational work that we have done supports the validity of the following conjecture, which is motivated by (6.2):

$$\frac{9}{8+r^2} < \frac{K(r)}{\log(4/\sqrt{1-r^2})} < \frac{9.2}{8+r^2}$$

for $0 < r < 1$.

(3) For $0 < r < 1$, the double inequality

$$1 + \frac{r^2(1-r^2)}{12-2r^2} < \frac{K(r)}{\log(4/\sqrt{1-r^2})} < \frac{10.2}{9+r^2}$$

is true. The upper bound is due to J. S. Frame [F] in private communication, and the lower bound is contained in the referee's report of this paper.

Acknowledgements. The authors wish to express appreciation to Mr. John Pemberton of the University of Auckland for the computer-drawn graphs in this paper. We thank the referee for bringing reference [CG] to our attention.

REFERENCES

[AS] M. ABRAMOWITZ AND I. A. STEGUN, eds., *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, Dover, New York, 1965.

[A] G. D. ANDERSON, *Derivatives of the conformal capacity of extremal rings*, Ann. Acad. Sci. Fenn. Ser. A I Math., 10 (1984), pp. 29-46.

[AV] G. D. ANDERSON AND M. K. VAMANAMURTHY, *Inequalities for elliptic integrals*, Publ. Inst. Math. (Beograd) (N.S.), 37 (1985), pp. 61-63.

[AVV1] G. D. ANDERSON, M. K. VAMANAMURTHY, AND M. VUORINEN, *Dimension-free quasiconformal distortion in n-space*, Trans. Amer. Math. Soc., 297 (1986), pp. 687-706.

[AVV2] ———, *Sharp distortion theorems for quasiconformal mappings*, Trans. Amer. Math. Soc., 305 (1988), pp. 95-111.

[AVV3] ———, *Special functions of quasiconformal theory*, Exposition. Math., 7 (1989), pp. 97-136.

[AVV4] ———, *Inequalities for the extremal distortion function*, Proc. Thirteenth Rolf Nevanlinna Colloquium, Joensuu, Finland, 1987, Lecture Notes in Math. 1351, Springer-Verlag, Berlin, New York, 1988, pp. 1-11.

[BB1] J. M. BORWEIN AND P. B. BORWEIN, *The arithmetic-geometric mean and fast computation of elementary functions*, SIAM Rev. 26 (1984), pp. 351-366.

[BB2] ———, *Pi and the* AGM, John Wiley, New York, 1987.

[Bo] F. BOWMAN, *Introduction to Elliptic Functions with Applications*, Dover, New York, 1961.

[BF] P. F. BYRD AND M. D. FRIEDMAN, *Handbook of Elliptic Integrals for Engineers and Physicists*, Grundlehren Math. Wiss. 57, Springer-Verlag, Berlin, New York, 1954.

[CG] B. C. CARLSON AND J. L. GUSTAFSON, *Asymptotic expansion of the first elliptic integral*, SIAM J. Math. Anal., 16 (1985), pp. 1072-1092.

[C] A. CAYLEY, *An Elementary Treatise on Elliptic Functions*, Deighton, Bell, and Co., Cambridge, U.K., 1876.

[E] A. ENNEPER, *Elliptische Functionen, Theorie und Geschichte*, Second edition, Louis Nebert, Halle, Germany, 1890.

[F] J. S. FRAME, Letter to G. D. Anderson, dated October 19, 20, 1988.

[Fr] C.-E. FRÖBERG, *Complete Elliptic Integrals*, CWK Gleerup, Lund, Sweden, 1957.

[G1] F. W. GEHRING, *Symmetrization of rings in space*, Trans. Amer. Math. Soc., 101 (1961), pp. 499-519.

[G2] ———, *Inequalities for condensers, hyperbolic capacity, and extremal lengths*, Michigan Math. J., 18 (1971), pp. 1-20.

[LF] A. V. LEBEDEV AND R. M. FEDOROVA, *A Guide to Mathematical Tables*, Pergamon Press, Elmsford, NY, 1960.

[LeVu] M. LEHTINEN AND M. VUORINEN, *On Teichmüller's modulus problem in the plane*, Rev. Roumaine Math. Pures Appl., 23 (1988), pp. 97-106.

[LV] O. LEHTO AND K. I. VIRTANEN, *Quasiconformal Mappings in the Plane*, Grundlehren Math. Wiss. 126, Second edition, Springer-Verlag, Berlin, New York, 1973.

[To] J. TODD, *Basic Numerical Mathematics*, Vol. 1, Birkhäuser-Verlag, Basel, Stuttgart, 1979.

[Tr] F. TRICOMI, *Lectures on the Use of Special Functions by Calculations with Electronic Computers*, Lecture Series 47, Institute for Fluid Dynamics and Applied Mathematics, University of Maryland, College Park, MD, 1966.

[Vu1] M. VUORINEN, *On Teichmüller's modulus problem in R^n*, Math. Scand., 63 (1988), pp. 101-119.

[Vu2] ———, *Conformal Geometry and Quasiregular Mappings*, Lecture Notes in Math. 1319, Springer-Verlag, Berlin, New York, 1988.

[WW] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Fourth edition, Cambridge University Press, Cambridge, London, 1958.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 551-576, May 1990
© 1990 Society for Industrial and Applied Mathematics

THE PLATE PARADOX FOR HARD AND SOFT SIMPLE SUPPORT*

I. BABUŠKA† AND J. PITKÄRANTA‡

Abstract. This paper studies the plate-bending problem with hard and soft simple support. It shows that in the case of hard support, the plate paradox, which is known to occur in the Kirchhoff model, is also present in the three-dimensional model and the Reissner-Mindlin model. The paradox consists of the fact that, on a sequence of convex polygonal domains converging to a circle, the solutions of the corresponding plate-bending problems with a fixed uniform load do not converge to the solution of the limit problem. The paper also shows that the paradox is not present when soft simple support is assumed. Some practical aspects are briefly discussed.

Key words. plates, Kirchhoff model, Reissner-Mindlin model, simply supported plate, plate paradox

AMS(MOS) subject classifications. 35J50, 35J55, 73C20, 73C99

1. Introduction. The Kirchhoff model of a plate is usually accepted as a good approximation to the three-dimensional model for thin plates. In the case of simply supported polygonal plates, however, the Kirchhoff model is known to suffer from unphysical phenomena that can lead to a large error of the model in some situations. In particular, the following paradox, referred to below as the plate paradox, occurs [2], [4]. Consider a sequence $\{\omega_n\}$ of convex polygonal domains approaching a circle. For each n, let w_n be the transverse deflection corresponding to the Kirchhoff model of the plate-bending problem, where the plate occupying the region ω_n is simply supported on $\partial\omega_n$ and is subject to a uniform load $p(x) \equiv 1$. Furthermore, let w_c be the solution to the limit problem, i.e., that on the circle. Then as $n \to \infty$, the sequence $\{w_n\}$ converges pointwise, but the limit w_∞ is different from w_c. For example, at the center of the circle the error of w_∞ is about 40 percent. Some other related plate paradoxes are given in [14], [15]. Practical implications occur, for example, in the finite-element method when the domain is approximated by a polygon with sidelength $h \to 0$. For further aspects see also [8], [18], [21], [23], and [25].

It is often assumed that the plate paradox is caused by the assumption of vanishing vertical shear strains that is implicit in the Kirchhoff model. This has been supported, e.g., by a note (see [3]) that the paradox is not present when the Reissner-Mindlin model instead of the Kirchhoff model is used. The aim of this paper is to locate the source of the paradox more precisely. We show that it is the way the *boundary conditions* are imposed in the Kirchhoff model that causes the paradox, and not the overall assumption of vanishing shear strains.

In the three-dimensional model of the plate, the boundary condition of simple support is typically imposed by requiring that the vertical component of the displacement (or at least its average in the vertical direction) vanish on the edge of the plate. On the other hand, the Kirchhoff model effectively imposes the more restrictive condition that all tangential displacements must vanish on the edge. Of course, it is

* Received by the editors January 27, 1989; accepted for publication April 11, 1989.

† Institute for Physical Science and Technology and Department of Mathematics, University of Maryland, College Park, Maryland 20742. The work of this author was partially supported by Office of Naval Research contract N00014-85-K0169 and by National Science Foundation grant DMS-85-16191.

‡ Department of Mathematics, University of Maryland, College Park, Maryland 20742 and Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo, Helsinki, Finland. The work of this author was supported by the Finnish Academy and by the Department of Mathematics, University of Maryland, College Park, Maryland.

also possible to impose such "hard" boundary conditions in other plate models, e.g., in the Reissner–Mindlin model (cf. [22]) or in the three-dimensional model itself. We show that in such a case, *the plate paradox occurs* in both the Reissner–Mindlin model and in the three-dimensional model. On the other hand, we also show that *the paradox does not occur* in these models in case of "soft" support, where only the vertical displacements are restricted on the edge of the plate. Hence, we are led to the conclusion that the paradox is caused by the hard boundary conditions that are intrinsic in the Kirchhoff model.

Our results are based on energy estimates relating the three-dimensional model and the Reissner–Mindlin model to the Kirchhoff model. Such estimates can be derived by combining the energy and complementary energy principles associated to the plate-bending problem. They were, in fact, applied early by Morgenstern [16], [17] to prove that the Kirchhoff model is the correct asymptotic limit of the three-dimensional model as the thickness of the plate tends to zero. Although the assumption of a smooth domain is implicit in Morgenstern's work, we can easily extend the analysis techniques of [16] to more general situations. In particular, we show here that in a sequence of convex polygonal domains converging to a circle, the relative error of the Kirchhoff model, when compared to the three-dimensional model with *hard* support, is uniformly of order $\mathcal{O}(h^{1/2})$ in the energy norm, where h is the thickness of the plate. Moreover, by similar techniques we show that the gap between Reissner–Mindlin and Kirchhoff models is uniformly of order $\mathcal{O}(h)$ under the same assumptions. Finally, we show that on a smooth domain, the three models are at most $\mathcal{O}(h^{1/2})$ apart. Hence we conclude that the plate paradox must occur in the hard-support models if h is fixed and sufficiently small.

Let us mention that our results are in parallel with recent benchmark calculations [7]. These calculations confirm, in particular, that the error of the Kirchhoff model with respect to the three-dimensional model is primarily due to the assumed hard boundary conditions on simply supported polygonal plates. For example, in the case of a uniformly loaded square plate of thickness $h = \text{side length}/100$, the relative error of the Kirchhoff model in energy norm is approximately 11 percent when compared to the three-dimensional model with soft support, and approximately 2 percent when compared to the hard-support model [7]. This example also shows that the error of the Kirchhoff model may be quite large even for relatively thin plates of simple shape.

The results above show that imposing various boundary conditions that are seemingly close, such as hard and soft simple support, can influence the solution in the entire domain and not only in the boundary layer. Very likely such effects also occur for other boundary conditions for both plates and shells. Therefore, since *any* boundary condition is an idealization of reality, finding the "correct" boundary conditions is an important and sometimes difficult part of building a dimensionally reduced model. For example, both soft and hard simple support can be poor approximations of the real "simple" support.

The plan of the paper is as follows. Section 2 gives the preliminaries and basic formulations of the plate problems. Section 3 elaborates on the variational formulations of the plate problems and presents various energy estimates. Section 4 addresses the problem of the plate paradox. Finally, Appendices A, B, and C present some auxiliary results needed in §§ 3 and 4.

2. Preliminaries. Consider an elastic plate of thickness h that occupies the region $\Omega = \omega \times (-h/2, h/2)$, where $\omega \in \mathbb{R}^2$ is a Lipschitz bounded domain. We assume that the plate is subject to given normal tractions p (i.e., the load) on $\omega \times \{-h/2\}$ and

$\omega \times \{h/2\}$ and that it is simply supported on $\partial\omega \times (-h/2, h/2)$ in such a way that if $\underline{u} = (u_1, u_2, u_3)$ is the displacement field, then

$$u_3(x) = 0, \qquad x \in \partial\omega \times \left(-\frac{h}{2}, \frac{h}{2}\right) \tag{2.1}$$

and the other two conditions are natural boundary conditions describing homogenous (zero) components of tractions. Later this condition will be called the soft simple support. If we assume for the moment that no other geometric boundary conditions other than (2.1) are imposed, the plate-bending problem can be formulated as follows. Find the displacement field $\underline{u}_0$ that minimizes the quadratic functional of the total energy

$$\begin{aligned} F(\underline{u}) = \frac{1}{2}\int_\Omega \left\{\lambda(\operatorname{div}\underline{u})^2 + \mu \sum_{i,j=1}^{3} [\varepsilon_{ij}(\underline{u})]^2\right\} dx_1\, dx_2\, dx_3 \\ - \int_\omega p \frac{1}{2}\left[u_3\left(\cdot, \frac{h}{2}\right) + u_3\left(\cdot, -\frac{h}{2}\right)\right] dx_1\, dx_2 \end{aligned} \tag{2.2}$$

in the Sobolev space $[H^1(\Omega)]^3$ under the boundary condition (2.1). Here $\underline{\varepsilon} = \{\varepsilon_{ij}\}_{i,j=1}^3$, $\varepsilon_{ij} = \frac{1}{2}((\partial u_i/\partial x_j) + (\partial u_j/\partial x_i))$ is the strain tensor, and λ and μ are the Lamé coefficients of the material, i.e.,

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \qquad \mu = \frac{E}{1+\nu},$$

where E is the Young modulus and ν is the Poisson ratio, $0 \leqq \nu \leqq \frac{1}{2}$. We also assume that the surface traction p is symmetrically distributed with regard to the planar surfaces of the plate, i.e., we consider a pure bending problem.

So far we have assumed a special model of simple support based on simple geometric constraint (2.1). Of course there are many other possibilities. Later we will discuss another model—of hard simple support—and will discuss the effects of these models of simple support on the solution.

It is well known that if $h/\operatorname{diam}(\omega)$ is small, the three-dimensional plate-bending problem can be formulated in various dimensionally reduced forms (see, e.g., [1], [11], [22]). Here we consider two representatives of such formulations that are used in practice: the Kirchhoff model and the Reissner-Mindlin model (cf. [22] and the references therein).

In general, when ω is fixed and $h \to 0$, the three-dimensional formulation and the dimensionally reduced models converge to the same limit, provided that the load p is appropriately scaled (see below). Hence for sufficiently thin plates the models give practically the same solutions. However, as will be seen later, what is "sufficiently thin" can depend strongly on ω, i.e., the convergence can be very slow in some situations.

In the Kirchhoff model, we approximate the three-dimensional solution as

$$\underline{u}_0(x_1, x_2, x_3) \cong \left(-x_3 \frac{\partial w_K}{\partial x_1}(x_1, x_2), -x_3 \frac{\partial w_K}{\partial x_2}(x_1, x_2), w_K(x_1, x_2)\right),$$

where w_K minimizes the energy

$$F_K(w) = \frac{1}{2}\int_\omega \left\{\nu(\Delta w)^2 + (1-\nu)\sum_{i,j=1}^{2}\left(\frac{\partial^2 w}{\partial x_i \partial x_j}\right)^2\right\} dx_1\, dx_2 - \int_\omega fw\, dx_1\, dx_2 \tag{2.3}$$

in the Sobolev space $H^2(\omega)$ under the boundary condition

$$w = 0 \quad \text{on } \partial\omega. \tag{2.4}$$

Here f is related to p as

$$f=\frac{p}{D}, \qquad D=\frac{Eh^3}{12(1-\nu^2)}. \tag{2.5}$$

When comparing different plate models with fixed ω and variable h we will assume below that f (and not p) is fixed. This ensures that the different models have the same (nontrivial) limit as $h\to 0$. For example, defining the average transverse deflection in the three-dimensional model as

$$w_0=\frac{1}{h}\int_{-h/2}^{h/2} u_3(\cdot, x_3)\,dx_3,$$

we can show (under fairly general assumptions on ω (see [10], [11], [16], [17], and § 3 below) that $\|w_0-w_K\|_{L_2(\omega)}\to 0$ as $h\to 0$.

In the Reissner–Mindlin model, we approximate $\underline{u}_0$ by

$$\underline{u}_0(x_1, x_2, x_3)\approx(-x_3\theta_{R,1}(x_1, x_2), -x_3\theta_{R,2}(x_1, x_2), w_R(x_1, x_2)),$$

where $(w_R, \underline{\theta}_R)$ minimizes the energy

$$\begin{aligned} F_R(w, \underline{\theta}) &= \frac{1}{2}\int_\omega \left\{\nu(\operatorname{div}\underline{\theta})^2+(1-\nu)\sum_{i,j=1}^{2}[\varepsilon_{ij}(\underline{\theta})]^2\right\} dx_1\,dx_2 \\ &\quad +\frac{1}{2}\left(\frac{\kappa}{h^2}\right)\int_\omega |\underline{\theta}-\underline{\nabla} w|^2\,dx_1\,dx_2-\int_\omega fw\,dx_1\,dx_2 \end{aligned} \tag{2.6}$$

in the Sobolev space $[H^1(\omega)]^3$ under the boundary condition (2.4). Here $\kappa=6(1-\nu)\kappa_0$, where $\kappa_0=O(1)$ is an additional shear correction factor that may take various values in practice.

We point out that the Kirchhoff approximation to $\underline{u}_0$ satisfies, in addition to (2.1), the boundary condition

$$(u_1t_1+u_2t_2)(x)=0, \qquad x\in\partial\omega\times\left(-\frac{h}{2}, \frac{h}{2}\right), \tag{2.7}$$

where $\underline{t}=(t_1, t_2)$ denotes the tangent to $\partial\omega$. This suggests that we should also consider the original plate-bending problem under such more restrictive geometric boundary conditions. Below we refer to the boundary conditions (2.1) and (2.7) and their counterpart in the Reissner–Mindlin model, i.e., (2.4) together with

$$\theta_1t_1+\theta_2t_2=0 \quad \text{on } \partial\omega, \tag{2.8}$$

as hard simple support in contrast to conditions (2.1) and (2.4), which we refer to as soft simple support. Hence, when using the Kirchhoff model we have in mind hard (not soft) simple support. Later we will show that the incapacity of the Kirchhoff model to represent soft simple support can be a severe deficiency of the model on polygonal domains.

3. Variational formulations of the plate-bending problem. Energy estimates. In §§ 3.1–3.3 and in the related Appendix A, we summarize first some basic characteristics of variational formalisms and energy principles associated to the plate-bending problem in its various forms. These results are basically known, but we present them here for the reader's convenience. In § 3.4 we prove some energy estimates relating the Kirchhoff model to both the Reissner–Mindlin model and the three-dimensional model, using the results of the previous subsections.

We assume that the plate occupies the region $\Omega = \omega \times (-h/2, h/2)$, where ω is a Lipschitz bounded domain. Our particular interest is in the cases where ω is either a convex polygon or a smooth domain.

We denote by $H^s(\omega)$, respectively, $H^s(\Omega)$, the usual Sobolev spaces with index $s > 0$. The seminorm and norm of the spaces $[H^s(\omega)]$ or $[H^s(\omega)]^k$ are denoted by $|\cdot|_{s,\omega}$ and $\|\cdot\|_{s,\omega}$, respectively, $|\cdot|_{s,\Omega}$ and $\|\cdot\|_{s,\Omega}$. By $(\cdot,\cdot)$ we mean the inner product of $[L_2(\omega)]^k$ or $[L_2(\Omega)]^k$, and by $\langle\cdot,\cdot\rangle$ the pairing of a space and its dual. The dual space of $H_0^1(\omega)$ will often be needed below and is denoted by $H^{-1}(\omega)$.

3.1. The three-dimensional model. Let us denote by N the space of horizontal rigid displacements of the plate

$$N = \{(\alpha_1 x_1 + \alpha_3 x_2, \alpha_2 x_2 - \alpha_3 x_1, 0),\ \alpha_i \in \mathbb{R},\ i = 1, 2, 3\}.$$

We define the space of (geometrically) admissible displacements in the case of soft simple support as

$$U = \left\{ \underline{u} \in [H^1(\Omega)]^3 \colon u_3 = 0 \text{ on } \partial\omega \times \left(-\frac{h}{2}, \frac{h}{2}\right),\ (\underline{u}, \underline{v}) = 0\ \ \forall \underline{v} \in N \right\} \tag{3.1a}$$

and in the case of hard simple support as

$$U = \left\{ \underline{u} \in [H^1(\Omega)]^3 \colon u_3 = t_1 u_1 + t_2 u_2 = 0 \text{ on } \partial\omega \times \left(-\frac{h}{2}, \frac{h}{2}\right),\ (\underline{u}, \underline{v}) = 0\ \ \forall \underline{v} \in N \right\}. \tag{3.1b}$$

(For simplicity, here we also remove all the horizontal rigid displacements in case of hard support.) Furthermore, we let $\mathcal{H}$ stand for the space of stress or strain tensors defined as

$$\mathcal{H} = \{\underline{\sigma} = (\sigma_{ij})_{i,j=1}^3 \colon \sigma_{ij} \in L_2(\Omega),\ \sigma_{ij} = \sigma_{ji}\},$$

and introduce a linear mapping $S \colon \mathcal{H} \to \mathcal{H}$ representing a scaled stress-strain relationship of a linear elastic material:

$$(S\underline{\tau})_{i,j} = D^{-1}[\lambda \operatorname{tr}(\underline{\tau})\delta_{ij} + \mu\tau_{ij}],$$

where λ and μ are the Lamé coefficients and the scaling factor D is as in (2.5). Then S is one to one and

$$(S^{-1}\underline{\tau})_{i,j} = \frac{D}{E}[(1+\nu)\tau_{ij} - \nu \operatorname{tr}(\underline{\tau})\delta_{ij}]. \tag{3.2}$$

Moreover, S and S^{-1} are self-adjoint if $\mathcal{H}$ is supplied with the natural inner product

$$(\underline{\sigma}, \underline{\tau})_{\mathcal{H}} = \sum_{i,j=1}^{3} (\sigma_{ij}, \tau_{ij}).$$

Let us further define the bilinear forms

$$\mathcal{A}(\underline{u}, \underline{v}) = (\underline{\varepsilon}(\underline{u}), S\underline{\varepsilon}(\underline{v}))_{\mathcal{H}}, \qquad \underline{u}, \underline{v} \in U,$$
$$\mathcal{B}(\underline{\sigma}, \underline{\tau}) = (\underline{\sigma}, S^{-1}\underline{\tau})_{\mathcal{H}}, \qquad \underline{\sigma}, \underline{\tau} \in \mathcal{H},$$

and the linear functional

$$Q(\underline{v}) = \frac{1}{2}\int_\omega f\left[v_3\left(\cdot, \frac{h}{2}\right) + v\left(\cdot, -\frac{h}{2}\right)\right] dx_1\, dx_2,$$

where it is assumed that $f \in L_2(\omega)$, to imply that Q is a bounded linear functional on U (by standard trace inequalities).

In the notation above, the *energy principle* states that the displacement field $\underline{u}_0$ due to the load $f=(p/D)\in L_2(\omega)$ is determined as the solution to the following minimization problem. Find $\underline{u}_0\in U$ that minimizes in U the functional

$$\mathcal{F}(\underline{u})=\tfrac{1}{2}\mathcal{A}(\underline{u},\underline{u})-Q(\underline{u}).$$

The existence and uniqueness of $\underline{u}_0$ is due to the following coercivity inequality, known as the Korn inequality (cf. [19]).

LEMMA 3.1. *If U is defined by* (3.1a), *there is a positive constant c such that*

$$\mathcal{A}(\underline{u},\underline{u})\geqq c\|\underline{u}\|^2_{1,\Omega},\qquad \underline{u}\in U. \tag{3.3}$$

We point out that the constant in (3.3) depends on ω (and h), although it is positive for any given Lipschitz domain. In Appendix B we show that the constant in (3.3) remains uniformly positive over a certain family of domains, a result needed in § 5 below.

Given f and the corresponding displacement field $\underline{u}_0$, let $\underline{\sigma}_0=S\underline{u}_0$ be the corresponding (scaled) stress field. The pair $(\underline{u},\underline{\sigma})=(\underline{u}_0,\underline{\sigma}_0)$ is then the solution to the following variational problem. Find $(\underline{u},\underline{\sigma})\in U\times\mathcal{H}$ such that

$$\mathcal{B}(\underline{\sigma},\underline{\tau})-(\underline{\varepsilon}(\underline{u}),\underline{\tau})_{\mathcal{H}}=0,\qquad \underline{\tau}\in\mathcal{H}, \tag{3.4a}$$

$$(\underline{\sigma},\underline{\varepsilon}(\underline{v}))=Q(\underline{v}),\qquad \underline{v}\in U. \tag{3.4b}$$

It can be easily verified following [5] and [9] (see Appendix A) that the solution to (3.4) exists and is unique.

Finally, we mention that according to the *complementary energy principle*, $\underline{\sigma}_0$ is found alternatively as the solution to the following minimization problem [19]. Find $\underline{\sigma}_0\in\mathcal{H}$ that minimizes in $\mathcal{H}$ the functional

$$\mathcal{G}(\underline{\sigma})=\tfrac{1}{2}\mathcal{B}(\underline{\sigma},\underline{\sigma})$$

under the constraint (3.4b).

In § 4 below we need the following corollary of the two energy principles (cf. [16]).

LEMMA 3.2. *For any $(\underline{u},\underline{\sigma})\in U\times\mathcal{H}$ such that $\underline{\sigma}$ satisfies* (3.4b), *the following identity holds*:

$$\tfrac{1}{2}\mathcal{A}(\underline{u}_0-\underline{u},\underline{u}_0-\underline{u})+\tfrac{1}{2}\mathcal{B}(\underline{\sigma}_0-\underline{\sigma},\underline{\sigma}_0-\underline{\sigma})=\mathcal{F}(\underline{u})+\mathcal{G}(\underline{\sigma}).$$

Proof. It follows from the energy principle that

$$\mathcal{A}(\underline{u}_0,\underline{v})=Q(\underline{v}),\qquad \underline{v}\in U,$$

and from the complementary energy principle that

$$\mathcal{B}(\underline{\sigma}_0,\underline{\tau})=0,\qquad \underline{\tau}\in\mathcal{H}\colon(\underline{\tau},\underline{\varepsilon}(\underline{v}))_{\mathcal{H}}=Q(\underline{v})\quad\forall\underline{v}\in U.$$

Therefore, in particular, $\mathcal{A}(\underline{u}_0,\underline{u})=Q(\underline{u})$ and $\mathcal{B}(\underline{\sigma}_0,\underline{\sigma})=\mathcal{B}(\underline{\sigma}_0,\underline{\sigma}_0)$, and hence

$$\begin{aligned}\tfrac{1}{2}\mathcal{A}(\underline{u}_0-\underline{u},\underline{u}_0-\underline{u})+\tfrac{1}{2}\mathcal{B}(\underline{\sigma}_0-\underline{\sigma},\underline{\sigma}_0-\underline{\sigma})&=[\tfrac{1}{2}\mathcal{A}(\underline{u},\underline{u})-\mathcal{A}(\underline{u}_0,\underline{u})+\tfrac{1}{2}\mathcal{B}(\underline{\sigma},\underline{\sigma})]\\&\quad+[\tfrac{1}{2}\mathcal{A}(\underline{u}_0,\underline{u}_0)+\tfrac{1}{2}\mathcal{B}(\underline{\sigma}_0,\underline{\sigma}_0)-\mathcal{B}(\underline{\sigma}_0,\underline{\sigma})]\\&=[\mathcal{F}(\underline{u})+\mathcal{G}(\underline{\sigma})]+[\tfrac{1}{2}\mathcal{A}(\underline{u}_0,\underline{u}_0)-\tfrac{1}{2}\mathcal{B}(\underline{\sigma}_0,\underline{\sigma}_0)]\\&=\mathcal{F}(\underline{u})+\mathcal{G}(\underline{\sigma}).\end{aligned}$$

□

3.2. The Reissner–Mindlin model. In the Reissner-Mindlin model geometrically admissible displacements $(w,\underline{\theta})$ span the space $H^1_0(\omega)\times V$, where either

$$V=[H^1(\omega)]^2 \tag{3.5a}$$

or

$$V=\{\underline{\theta}\in[H^1(\omega)]^2:\ t_1\theta_1+t_2\theta_2=0 \text{ on } \partial\omega\} \tag{3.5b}$$

corresponding to soft and hard boundary conditions, respectively. We let $\mathcal{K}$ stand for the space of momentum and curvature tensors:

$$\mathcal{K}=\{\underline{m}=(m_{ij})_{i,j=1}^2,\ m_{ij}\in L_2(\omega),\ m_{12}=m_{21}\},$$

and supply $\mathcal{K}$ with the natural inner product

$$(\underline{m},\underline{k})_{\mathcal{K}}=\sum_{i,j=1}^{2}(m_{ij},k_{ij}).$$

Furthermore, we introduce the linear mapping $T:\mathcal{K}\to\mathcal{K}$ as defined by

$$(T\underline{k})_{ij}=\nu\,\mathrm{tr}\,(\underline{k})\delta_{ij}+(1-\nu)k_{ij},\qquad \underline{k}\in\mathcal{K}.$$

The inverse of T is given by

$$(T^{-1}\underline{k})_{ij}=\frac{1}{1-\nu}k_{ij}-\frac{\nu}{1+\nu}\,\mathrm{tr}\,(\underline{k})\delta_{ij}, \tag{3.6}$$

and obviously T and T^{-1} are self-adjoint.

Finally we introduce the bilinear forms

$$\mathcal{A}_R(w,\underline{\theta};z,\underline{\varphi})=(\underline{\varepsilon}(\underline{\theta}),T\underline{\varepsilon}(\underline{\varphi}))+\left(\frac{\kappa}{h^2}\right)(\underline{\theta}-\underline{\nabla}w,\underline{\varphi}-\underline{\nabla}z),\quad w,z\in H_0^1(\omega),\quad \underline{\theta},\underline{\varphi}\in V,$$

$$\mathcal{B}_R(\underline{m},\underline{\gamma};\underline{k},\underline{\zeta})=(\underline{m},T^{-1}\underline{k})_{\mathcal{K}}+\left(\frac{h^2}{\kappa}\right)(\underline{\gamma},\underline{\zeta}),\quad \underline{m},\underline{k}\in\mathcal{K},\quad \underline{\gamma},\underline{\zeta}\in[L_2(\omega)]^2,$$

where $\underline{\varepsilon}(\underline{\theta})=(\varepsilon_{ij}(\underline{\theta}))_{i,j=1}^2$ and κ are as in (2.6).

In the notation above, the Reissner–Mindlin formulation of the plate-bending problem, as stated according to the energy principle, is to find the pair $(w_R,\underline{\theta}_R)\in H_0^1(\omega)\times V$ that minimizes in $H_0^1(\omega)\times V$ the functional

$$\mathcal{F}_R(w,\underline{\theta})=\tfrac{1}{2}\mathcal{A}_R(w,\underline{\theta};w,\underline{\theta})-\langle f,w\rangle$$

for a given $f\in H^{-1}(\omega)$.

The existence and uniqueness of $(w_R,\underline{\theta}_R)$ is the consequence of the following lemma, which is proved in Appendix B in slightly more general form (see Lemma B.2 of Appendix B).

LEMMA 3.3. *There is a positive constant c such that*

$$(\underline{\varepsilon}(\underline{\theta}),T\underline{\varepsilon}(\underline{\theta}))_{\mathcal{K}}+\|\underline{\theta}-\underline{\nabla}w\|_{0,\omega}^2\geqq c(\|\underline{\theta}\|_{1,\omega}^2+\|w\|_{1,\omega}^2),$$
$$\underline{\theta}\in[H^1(\omega)]^2,\quad w\in H_0^1(\omega).$$

Remark 3.1. Regarding the validity of Lemma 3.3 uniformly over a sequence of domains, see Appendix B. (Such a result is needed in § 5 below.)

The analogy of the variational formulation (3.4) is stated for the Reissner–Mindlin model as follows. Find $(w,\underline{\theta},\underline{m},\underline{\gamma})\in H_0^1(\omega)\times V\times\mathcal{K}\times[L_2(\omega)]^2$ such that

$$(\underline{m},T^{-1}\underline{k})_{\mathcal{K}}-(\underline{\varepsilon}(\underline{\theta}),\underline{k})_{\mathcal{K}}=0,\qquad \underline{k}\in\mathcal{K}, \tag{3.7a}$$

$$(h^2/\kappa)(\underline{\gamma},\underline{\zeta})-(\underline{\theta}-\underline{\nabla}w,\zeta)=0,\qquad \underline{\zeta}\in[L_2(\omega)]^2, \tag{3.7b}$$

$$(\underline{m},\underline{\varepsilon}(\underline{\varphi}))_{\mathcal{K}}+(\underline{\gamma},\underline{\varphi})=0,\qquad \underline{\varphi}\in V, \tag{3.7c}$$

$$-(\underline{\gamma},\underline{\nabla}z)=\langle f,z\rangle,\qquad z\in H_0^1(\omega). \tag{3.7d}$$

The (unique; see Appendix A) solution to this problem is $(w_R, \underline{\theta}_R, \underline{\underline{m}}_R, \underline{\gamma}_R)$, where $\underline{\underline{m}}_R = T\underline{\underline{\varepsilon}}(\underline{\theta}_R)$ and $\underline{\gamma}_R = (\kappa/h^2)(\underline{\theta}_R - \underline{\nabla} w_R)$ have the physical meaning of momentum and (vertical) shear stress field, respectively, both being scaled by a factor D^{-1}.

Finally we note that the pair $(\underline{\underline{m}}_R, \underline{\gamma}_R)$ can be obtained alternatively as the solution to the following minimization problem (the complementary energy principle). Find $(\underline{\underline{m}}_R, \underline{\gamma}_R) \in \mathcal{H} \times [L_2(\omega)]^2$ that minimizes in $\mathcal{H} \times [L_2(\omega)]^2$ the functional

$$\mathcal{G}_R(\underline{\underline{m}}, \underline{\gamma}) = \tfrac{1}{2}\mathcal{B}_R(\underline{\underline{m}}, \underline{\gamma}; \underline{\underline{m}}, \underline{\gamma})$$

under the constraints (3.7c, d).

Upon combining the two energy principles we obtain, in analogy with Lemma 3.2, the following lemma.

LEMMA 3.4. *For any* $(w, \theta) \in H_0^1(\omega) \times V$ *and for any* $(\underline{\underline{m}}, \underline{\gamma}) \in \mathcal{H} \times [L_2(\omega)]^2$ *satisfying* (3.7c, d) *the following identity holds*:

$$\tfrac{1}{2}\mathcal{A}_R(w_R - w, \underline{\theta}_R - \underline{\theta}; w_R - w, \underline{\theta}_R - \underline{\theta}) + \tfrac{1}{2}\mathcal{B}_R(\underline{\underline{m}}_R - \underline{\underline{m}}, \underline{\gamma}_R - \underline{\gamma}; \underline{\underline{m}}_R - \underline{\underline{m}}, \underline{\gamma}_R - \underline{\gamma}) = \mathcal{F}_R(w, \underline{\theta}) + \mathcal{G}_R(\underline{\underline{m}}, \underline{\gamma}).$$

3.3. The Kirchhoff model. Upon introducing the space

$$W = \{z \in H^2(\omega): z = 0 \text{ on } \partial\omega\},$$

we formulate the plate-bending problem according to the Kirchhoff model as follows. Given $f \in W'$ (=dual space of W), find $w_K \in W$ that minimizes in W the energy functional

$$\mathcal{F}_K(w) = \tfrac{1}{2}(\underline{\underline{\varepsilon}}(\underline{\nabla} w), T\underline{\underline{\varepsilon}}(\underline{\nabla} w))_{\mathcal{H}} - \langle f, w\rangle,$$

where T and $(\cdot, \cdot)_{\mathcal{H}}$ are the same as in the Reissner–Mindlin model. The existence and uniqueness of $w_{\mathcal{H}}$ in the consequence of the coercivity inequality

$$(\underline{\underline{\varepsilon}}, (\underline{\nabla} w), T\underline{\underline{\varepsilon}}(\underline{\nabla} w))_{\mathcal{H}} \geqq c\|w\|_{2,\omega}^2, \qquad w \in W,$$

which itself is an easy consequence of Lemma 3.3. Note that $w_{\mathcal{H}}$ is uniquely defined, in particular, if $f \in H^{-1}(\omega)$, and note also that the pair $(w_K, \underline{\theta}_K)$, where $\underline{\theta}_K = \underline{\nabla} w_K$, minimizes the Reissner functional $\mathcal{F}_R$ over the subspace $Z \subset H_0^1(\omega) \times V$ defined by

$$Z = \{(w, \underline{\theta}) \in W \times V: \theta = \underline{\nabla} w\}.$$

For the Kirchhoff model, the analogy of the mixed variational formulation (3.7) is the following. Given $f \in W'$, find $(w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}) \in W \times V \times \mathcal{H} \times V'$ (where V' is the dual space of V) such that

$$(\underline{\underline{m}}, T^{-1}\underline{\underline{k}})_{\mathcal{H}} - (\underline{\underline{\varepsilon}}(\underline{\theta}), \underline{\underline{k}})_{\mathcal{H}} = 0, \qquad \underline{\underline{k}} \in \mathcal{H}, \tag{3.8a}$$

$$\langle \underline{\theta} - \underline{\nabla} w, \underline{\zeta}\rangle = 0, \qquad \underline{\zeta} \in V', \tag{3.8b}$$

$$(\underline{\underline{m}}, \underline{\underline{\varepsilon}}(\underline{\varphi}))_{\mathcal{H}} + \langle \underline{\gamma}, \underline{\varphi}\rangle = 0, \qquad \underline{\varphi} \in V, \tag{3.8c}$$

$$-\langle \underline{\gamma}, \underline{\nabla} z\rangle = \langle f, z\rangle, \qquad z \in W. \tag{3.8d}$$

LEMMA 3.5. *The variational problem* (3.8) *is well posed and the unique solution is* $(w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}) = (w_K, \underline{\theta}_K, \underline{\underline{m}}_K, \underline{\gamma}_K)$, *where* $\underline{\theta}_K = \underline{\nabla} w_K$, $\underline{\underline{m}}_K = T\underline{\underline{\varepsilon}}(\underline{\theta}_K)$, *and* $\underline{\gamma}_K$ *is defined by* (3.8c), *i.e.*,

$$\langle \underline{\gamma}_K, \underline{\varphi}\rangle = -(\underline{\underline{m}}_K, \underline{\underline{\varepsilon}}(\underline{\varphi}))_{\mathcal{H}}, \qquad \underline{\varphi} \in V. \tag{3.9}$$

Proof. If $(w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}) = (w_K, \underline{\theta}_K, \underline{\underline{m}}_K, \underline{\gamma}_K)$, (3.8a–c) hold trivially. Moreover, since w_K minimizes $\mathcal{F}_K$ in W, we have $(\underline{\underline{m}}_K, \underline{\underline{\varepsilon}}(\underline{\nabla} z))_{\mathcal{H}} = (\underline{\underline{\varepsilon}}(\underline{\nabla} w_K), T\underline{\underline{\varepsilon}}(\underline{\nabla} z))_{\mathcal{H}} = \langle f, z\rangle$ for all $z \in W$; so, by (3.9), (3.8d) holds as well. The well-posedness is proved in Appendix A. □

Remark 3.2. Note that although w_K, $\underline{\theta}_K$, and $\underline{\underline{m}}_K$ obviously do not depend on the way the space V is defined in (3.5), $\underline{\gamma}_K$ certainly does (see below). Hence in this (somewhat weak) sense the "soft" and "hard" formulations are still separate even in the Kirchhoff model.

We need below the following specific result related to the case where ω is a convex polygon.

LEMMA 3.6. *Let w_K be defined as above assuming that ω is a convex polygon and that $f \in H^{-1}(\omega)$. Furthermore, let $\rho \in H_0^1(\omega)$ and $\psi \in H_0^1(\omega)$ be such that*

$$(\underline{\underline{\nabla}}\rho, \underline{\underline{\nabla}}\xi) = (\psi, \xi), \qquad \xi \in H_0^1(\omega), \tag{3.10a}$$

$$(\underline{\underline{\nabla}}\psi, \underline{\underline{\nabla}}\xi) = \langle f, \xi\rangle, \qquad \xi \in H_0^1(\omega). \tag{3.10b}$$

Then $\rho = w_K$ and $\psi = -\Delta w_K$.

Proof. From (3.10a, b) it is obvious that $\psi = -\Delta\rho \in H_0^1(\omega)$, so it suffices to show that $\rho = w_K$. First, since $\psi \in H^1(\omega)$ and since ω is a convex polygon, it follows from (3.10a) that $\rho \in H^2(\omega)$ and $\rho \in H^3(\tilde{\omega})$, $\tilde{\bar{\omega}} \subset \bar{\omega} - UA_i$, A_i being the vertices of ω; i.e., $\rho \in W$ (cf. [13]). Moreover, since $\rho = \Delta\rho = 0$ almost everywhere on $\partial\omega$ and since $\partial\omega$ consists of straight-line segments only, it follows that $\partial^2\rho/\partial t^2 = \partial^2\rho/\partial n^2 = 0$ almost everywhere on $\partial\omega$. Therefore, and noting also that $\partial z/\partial t = 0$ almost everywhere on $\partial\omega$ if $z \in W$, integrating by parts shows that

$$\begin{aligned}(\underline{\underline{\nabla}}\psi, \underline{\underline{\nabla}}z) &= -(\underline{\underline{\nabla}}(\Delta\rho), \underline{\underline{\nabla}}z) = -\nu(\underline{\underline{\nabla}}(\Delta\rho), \underline{\underline{\nabla}}z) - (1-\nu)\sum_{i,j=1}^{2}\left(\frac{\partial^3\rho}{\partial x_i \partial^2 x_j^2}, \frac{\partial z}{\partial x_i}\right)\\ &= (\underline{\underline{\varepsilon}}(\underline{\underline{\nabla}}\rho), T\underline{\underline{\varepsilon}}(\underline{\underline{\nabla}}z)) - \int_{\partial\omega}\left[\nu\Delta\rho + (1-\nu)\frac{\partial^2\rho}{\partial n^2}\right]\frac{\partial z}{\partial n}\,ds\\ &= (\underline{\underline{\varepsilon}}(\underline{\underline{\nabla}}\rho), T\underline{\underline{\varepsilon}}(\underline{\underline{\nabla}}z)), \qquad z \in W.\end{aligned}$$

Hence, by (3.10b), $(\underline{\underline{\varepsilon}}(\underline{\underline{\nabla}}\rho), T\underline{\underline{\varepsilon}}(\nabla z)) = \langle f, z\rangle$, for all $z \in W$, so ρ minimizes $\mathscr{F}_K$ in W and accordingly, $\rho = w_K$. □

We can now prove the following result that will be needed in the next subsection.

LEMMA 3.7. *Let ω be either a convex polygon or a smooth domain, and let $(w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}) = (w_K, \underline{\theta}_K, \underline{\underline{m}}_K, \underline{\gamma}_K) \in W \times V \times \mathscr{K} \times V'$ be the solution to (3.8) for a given $f \in H^{-1}(\omega)$, and with V defined by (3.5b). Then $\underline{\gamma}_K = -\underline{\underline{\nabla}}(\Delta w_K) \in [L_2(\omega)]^2$ and $(w_K, \underline{\theta}, \underline{\underline{m}}_K, \underline{\gamma}_K)$ is a solution to (3.7) with $h = 0$ in (3.7b). Moreover, if ω is a convex polygon, then $\|\underline{\gamma}_K\|_{0,\omega} = \|f\|_{-1,\omega}$, where*

$$\|f\|_{-1,\omega} = \sup_{z \in H_0^1(\omega)} \frac{\langle f, z\rangle}{|z|_{1,\omega}},$$

and if ω is a smooth domain, then $\|\underline{\gamma}_K\|_{0,\omega} \leqq C\|f\|_{-1,\omega}$, where C depends on ω.

Proof. If $\underline{\gamma}_K \in [L_2(\omega)]^2$ and $f \in H^{-1}(\omega)$, it follows from a simple closure argument that (3.8) remains valid if W is replaced by V and if $\langle\cdot,\cdot\rangle$ on the left side is replaced by $(\cdot,\cdot)$. To prove that $\underline{\gamma}_K = -\underline{\underline{\nabla}}(\Delta w_K)$, we integrate by parts in (3.9) to obtain

$$\begin{aligned}\langle\underline{\gamma}_K, \underline{\varphi}\rangle &= -\int_\omega \underline{\underline{\nabla}}(\Delta w_K)\cdot\underline{\varphi}\,dx_1\,dx_2 + \int_{\partial\omega}\left[\nu\Delta w_K + (1-\nu)\frac{\partial^2 w_K}{\partial n^2}\right]\underline{\varphi}\cdot\underline{n}\,ds\\ &\quad + \int_{\partial\omega}(1-\nu)\frac{\partial^2 w_K}{\partial n \partial t}\underline{\varphi}\cdot\underline{t}\,ds, \qquad \underline{\varphi} \in V.\end{aligned}$$

Here the first boundary integral vanishes because $\nu\Delta w + (1-\nu)(\partial^2 w/\partial n^2) = 0$ on $\partial\omega$ is the natural boundary condition associated to the problem of minimizing $\mathscr{F}_K$, and the

second boundary integral vanishes since $\underline{\varphi} \cdot \underline{t} = 0$, $\underline{\varphi} \in V$, assuming that V is defined by (3.5b). Hence $\underline{\gamma}_K = -\underline{\nabla}(\Delta w_K)$. On the other hand, from (3.10) we have $\underline{\gamma}_K \in [L_2(\omega)]^2$, and from the well-posedness of (3.8) we see that indeed $\underline{\gamma}_K = -\underline{\nabla}(\Delta w_K)$.

Having verified that $\underline{\gamma}_K = -\underline{\nabla}(\Delta w_K)$, we conclude from Lemma 3.6 that $\underline{\gamma}_K = \underline{\nabla}\psi$, where $\psi \in H_0^1(\omega)$ satisfies (3.10b), so $\|\underline{\gamma}_K\|_{0,\omega} = \|f\|_{-1,\omega}$ as asserted. Finally, if ω is a smooth domain, a standard elliptic regularity estimate implies that $\|\underline{\gamma}_K\|_{0,\omega} \leqq C\|w\|_{3,\omega} \leqq C_1\|f\|_{-1,\omega}$. □

Remark 3.3. It is essential for our results in the next section that when ω is a convex polygon, $\|\underline{\gamma}_K\|_{0,\omega}$ is bounded by $\|f\|_{-1,\omega}$ independently of ω, in contrast to the smooth domain, where the constant depends on ω.

3.4. Energy estimates in case of hard support. Let us define the energy norms

$$|||\underline{u}, \underline{\sigma}|||^2 = \mathcal{A}(\underline{u}, \underline{u}) + \mathcal{B}(\underline{\sigma}, \underline{\sigma}), \qquad (\underline{u}, \underline{\sigma}) \in U \times \mathcal{H},$$

$$|||w, \underline{\theta}, \underline{m}, \underline{\gamma}|||_R^2 = \mathcal{A}_R(w, \underline{\theta}; w, \underline{\theta}) + \mathcal{B}_R(\underline{m}, \underline{\gamma}; \underline{m}, \underline{\gamma}),$$

$$(w, \underline{\theta}, \underline{m}, \underline{\gamma}) \in H_0^1(\omega) \times V \times \mathcal{K} \times [L_2(\omega)]^2,$$

where the bilinear forms are as defined in §§ 3.1 and 3.2. Then by Lemma 3.2 we have the identity

$$|||\underline{u}_0 - \underline{u}, \underline{\sigma}_0 - \underline{\sigma}|||^2 = |||\underline{u}, \underline{\sigma}|||^2 - 2Q(\underline{u}) \tag{3.11}$$

whenever $\underline{u} \in U$ and $\underline{\sigma} \in \mathcal{H}$ satisfies the constraint (3.4b). Similarly, by Lemma 3.4,

$$|||w_R - w, \underline{\theta}_R - \underline{\theta}, \underline{m}_R - \underline{m}, \underline{\gamma}_R - \underline{\gamma}|||_R^2 = |||w, \underline{\theta}, \underline{m}, \underline{\gamma}|||_R^2 - 2\langle f, w\rangle, \tag{3.12}$$

where $(w, \underline{\theta}) \in H_0^1(\omega) \times V$ and $(\underline{m}, \underline{\gamma}) \in \mathcal{K} \times [L_2(\omega)]^2$ satisfies constraints (3.7c, d).

Let us first apply (3.12) to estimate the gap between the Reissner "quadruple" $(w_R, \underline{\theta}_R, \underline{m}_R, \underline{\gamma}_R)$ and the Kirchhoff "quadruple" $(w_K, \underline{\theta}_K, \underline{m}_K, \underline{\gamma}_K)$. By Lemma 3.7, the choice $(w, \underline{\theta}, \underline{m}, \underline{\gamma}) = (w_K, \underline{\theta}_K, \underline{m}_K, \underline{\gamma}_K)$ is legitimate in (3.12) under the assumptions that ω is either a convex polygon or a smooth domain; $f \in H^{-1}(\omega)$; and V is defined by (3.5b), i.e., the case of hard support. Upon simplifying the right side of (3.12), in this case we obtain the identity

$$|||w_R - w_K, \underline{\theta}_R - \underline{\theta}_K, \underline{m}_R - \underline{m}_K, \underline{\gamma}_R - \underline{\gamma}_K|||_R^2 = (h^2/\kappa)\|\underline{\gamma}_K\|_{0,\omega}^2,$$

which together with Lemma 3.7 leads to the following theorem.

THEOREM 3.1. *Let ω be either* (a) *a convex polygon or* (b) *a smooth domain, let $f \in H^{-1}(\omega)$, and let $(w_R, \underline{\theta}_R, \underline{m}_R, \underline{\gamma}_R)$ and $(w_K, \underline{\theta}_K, \underline{m}_K, \underline{\gamma}_K)$ be the solution to* (3.7) *and* (3.8), *respectively, where V is defined by* (3.5b). *Then in case* (a) *we have the identity*

$$|||w_R - w_K, \underline{\theta}_R - \underline{\theta}_K, \underline{m}_R - \underline{m}_K, \underline{\gamma}_R - \underline{\gamma}_K|||_R^2 = (h^2/\kappa)\|f\|_{-1,\omega}^2,$$

where $\|f\|_{-1,\omega}$ is defined as in Lemma 3.7, *and in case* (b) *the estimate*

$$|||w_R - w_K, \underline{\theta}_R - \underline{\theta}_K, \underline{m}_R - \underline{m}_K, \underline{\gamma}_R - \underline{\gamma}_K|||_R^2 = C(h^2/\kappa)\|f\|_{-1,\omega}^2,$$

where C depends on ω.

Remark 3.4. It is easy to verify that

$$|||w_R - w_K, \underline{\theta}_R - \underline{\theta}_K, \underline{m}_R - \underline{m}_K, \underline{\gamma}_R - \underline{\gamma}_K|||_R^2 \geqq E_K - E_R,$$

where E_R and E_K stand for the total energy of the plate in the Kirchhoff and Reissner–Mindlin models, respectively, i.e.,

$$E_K = \mathcal{F}_R(w_K, \underline{\theta}_K) = -\tfrac{1}{2}\langle f, w_K\rangle, \qquad E_R = \mathcal{F}_R(w_R, \underline{\theta}_R) = -\tfrac{1}{2}\langle f, w_R\rangle.$$

In particular, if ω is a convex polygon, Theorem 3.1 and Lemma 3.6 lead to the relative estimate

$$(E_K - E_R)/E_K \leqq C(\omega, f, \nu)h^2/\kappa_0,$$

where κ_0 is the shear correction factor, and

$$C(\omega, f, \nu) = -\frac{1}{6(1-\nu)} \frac{\int_\omega \Delta w_K f \, dx_1 \, dx_2}{\int_\omega w_K f \, dx_1 \, dx_2}.$$

For example, if ω is the unit square and $f(x) \equiv 1$, then $C(\omega, f, \nu) = 3.440428/(1-\nu)$.

Remark 3.5. In case of soft boundary conditions, constraint (3.7c) is more restrictive and rules out the choice $(\underline{\underline{m}}, \underline{\gamma}) = (\underline{\underline{m}}_K, \underline{\gamma}_K)$ in (3.12). It is still possible to find $(\tilde{\underline{\underline{m}}}_K, \tilde{\underline{\gamma}}_K) \in \mathcal{H} \times [L_2(\omega)]^2$, which is close to $(\underline{\underline{m}}_K, \underline{\gamma}_K)$ away from the boundary and satisfies all the required constraints [16]. With such a construction, it is possible to show that if both f and ω are sufficiently smooth, then

$$|||w_R - w_K, \underline{\theta}_R - \underline{\theta}_K, \underline{\underline{m}}_R - \tilde{\underline{\underline{m}}}_K, \underline{\gamma}_R - \tilde{\underline{\gamma}}_K|||_R^2 \leqq C(\omega, f)h.$$

For other estimates of this type see also [11], the references therein, and [20].

Next, we apply (3.11) to bound the difference between the three-dimensional solution and the Kirchhoff solution. To this end, we need to construct a three-dimensional extension $(\underline{u}_K, \underline{\underline{\sigma}}_K) \in U \times \mathcal{H}$ of the Kirchhoff solution $(w_K, \underline{\theta}_K, \underline{\underline{m}}_K, \underline{\gamma}_K)$. Following [16] we define $\underline{u}_K \in U$ as

$$\underline{u}_K = (-x_3\theta_{K,1}, -x_3\theta_{K,2}, w_K + \tfrac{1}{2}x_3^2\psi), \tag{3.13}$$

and $\underline{\underline{\sigma}}_K \in \mathcal{H}$ as

$$\begin{aligned} &\sigma_{K,ij} = -\alpha x_3 m_{K,ij}, && i,j = 1,2, \\ &\sigma_{K,i3} = \alpha(\tfrac{1}{2}x_3^2 - \tfrac{1}{8}h^2)\gamma_{K,i}, && i = 1,2, \\ &\sigma_{K,33} = \alpha(-\tfrac{1}{6}x_3^3 + \tfrac{1}{8}h^2 x_3)f, \end{aligned} \tag{3.14}$$

where $\alpha = 12/h^3$ and $\psi \in H_0^1(\omega)$ is so far unspecified. It is easy to check that $\underline{\underline{\sigma}}_K$ satisfies (3.4b) as far as U is defined by (3.1b), so (3.11) applies with the choice $(\underline{u}, \underline{\underline{\sigma}}) = (\underline{u}_K, \underline{\underline{\sigma}}_K)$ in this case. After a short computation, the right side of (3.11) can then be expressed as

$$\begin{aligned} |||\underline{u}_K, \underline{\underline{\sigma}}_K|||^2 - 2Q(\underline{u}_K) = &\frac{(1-\nu)^2}{1-2\nu} \int_\omega \left(\psi + \frac{\nu}{1-\nu}\Delta w_K\right)^2 dx_1 \, dx_2 \\ &+ \frac{3(1-\nu)}{160} h^2 \int_\omega |\nabla\psi|^2 \, dx_1 \, dx_2 \\ &+ \frac{1}{5(1-\nu)} h^2 \int_\omega (|\underline{\gamma}_K|^2 + \nu \Delta w_K f) \, dx_1 \, dx_2 \\ &+ \frac{17}{1680(1-\nu^2)} h^4 \int_\omega f^2 \, dx_1 \, dx_2 - \frac{1}{4} h^2 \int_\omega \psi f \, dx_1 \, dx_2. \end{aligned}$$

Now if ω is a convex polygon, the choice $\psi = (\nu/(1-\nu))\Delta w_K$ is legitimate and leads—recall also that $\|\gamma_K\|_{0,\omega}^2 = -\int_\omega \Delta w_K f \, dx_1 \, dx_2 = \|f\|_{-1,\omega}^2$ (see Lemmas 3.6 and 3.7)—to the identity

$$|||\underline{u}_K, \underline{\underline{\sigma}}_K|||^2 - 2Q(\underline{u}_K) = \frac{32 + 8\nu + 3\nu^2}{160(1-\nu)} h^2 \|f\|_{-1,\omega}^2 + \frac{17}{1680(1-\nu^2)} h^4 \int_\omega f^2 \, dx_1 \, dx_2.$$

On the other hand, if ω is a smooth domain, we can still find for any $\delta > 0$ a $\psi \in H_0^1(\omega)$ so that

$$\int_\omega \left(\psi - \frac{\nu}{1-\nu}\Delta w_K\right)^2 dx_1\, dx_2 \leqq C\delta\nu^2 \|\Delta w_K\|_{1,\omega}^2, \tag{3.15a}$$

$$\int_\omega |\nabla\psi|^2\, dx_1\, dx_2 \leqq C\delta^{-1}\nu^2 \|\Delta w_K\|_{1,\omega}^2. \tag{3.15b}$$

Since $\|\Delta w_K\|_{1,\omega} \leqq C(\omega)\|f\|_{-1,\omega}$, we obtain in this case, choosing $\delta = \sqrt{1-2\nu}\, h$, the estimate

$$|||\underline{u}_K, \underline{\underline{\varrho}}_K|||^2 - 2Q(\underline{u}_K) \leqq C(\omega)\frac{\nu^2 h^2}{\sqrt{1-2\nu}}\|f\|_{-1,\omega}^2 + \frac{17h^4}{1680(1-\nu^2)}\|f\|_{0,\omega}^2.$$

We thus conclude the following theorem.

THEOREM 3.2. *Assume that ω is either* (a) *a convex polygon or* (b) *a smooth domain. Let $f \in L_2(\omega)$, let $(\underline{u}_0, \underline{\underline{\varrho}}_0) \in U \times \mathcal{H}$ be the solution to* (3.4) *with U defined by* (3.1b), *and let $(\underline{u}_K, \underline{\underline{\varrho}}_K)$ be defined by* (3.13)-(3.14), *where $(w_K, \underline{\theta}_K, \underline{\underline{m}}_K, \underline{\gamma}_K) \in W \times V \times \mathcal{H} \times V'$ is the solution to* (3.8) *with V defined by* (3.5b), *and either $\psi = (\nu/(1-\nu))\Delta w_K$ (case* (a)) *or ψ satisfies* (3.15a, b) *with $\delta = \sqrt{1-2\nu}\, h$ (case* (b)). *Then in case* (a) *we have the identity*

$$|||\underline{u}_0 - \underline{u}_K, \underline{\underline{\varrho}}_0 - \underline{\underline{\varrho}}_K|||^2 = C_1(\nu)h^2\|f\|_{-1,\omega}^2 + C_2(\nu)h^4\|f\|_{0,\omega}^2$$

and in case (b) *the estimate*

$$|||\underline{u}_0 - \underline{u}_K, \underline{\underline{\varrho}}_0 - \underline{\underline{\varrho}}_K|||^2 \leqq C(\omega)[C_3(\nu)h + h^2]\|f\|_{-1,\omega}^2 + C_2(\nu)h^4\|f\|_{0,\omega}^2,$$

where $\|f\|_{-1,\omega}$ is defined as in Lemma 3.7 *and*

$$C_1(\nu) = \frac{32 + 8\nu + 3\nu^2}{160(1-\nu)}, \quad C_2(\nu) = \frac{17}{1680(1-\nu^2)}, \quad C_3(\nu) = \frac{\nu^2}{\sqrt{1-2\nu}}.$$

Remark 3.6. In the case of soft boundary conditions it is possible to show that, if ω is smooth and f is sufficiently smooth, then

$$|||\underline{u}_0 - \underline{u}_K, \underline{\underline{\varrho}}_0 - \underline{\underline{\tilde{\varrho}}}_K|||^2 \leqq C(\omega, f)[1 + C_3(\nu)]h,$$

where $\underline{\underline{\tilde{\varrho}}}_K$ is close to $\underline{\underline{\varrho}}_K$ away from the boundary strip $\partial\omega \times (-h/2, h/2)$ [11], [16].

4. The plate paradox. Let $\omega_0 \subset \mathbb{R}^2$ be the unit circular domain with the center at the origin, i.e.,

$$\omega^{[0]} = \{(x_1, x_2): r^2 = x_1^2 + x_2^2 < 1\}.$$

Furthermore, let $\omega^{[n]}$, $n = 1, 2, \cdots$, be the sequence of regular $(n+3)$-polygons such that

$$\bar{\omega}^{[n]} \subset \omega^{[n+1]} \subset \bar{\omega}^{[n+1]} \subset \omega^{[0]},$$

$$\omega^{[n]} \to \omega^{[0]} \quad \text{as } n \to \infty$$

in the sense that for any $x \in \omega^{[0]}$ there is $n(x) > 0$ such that $x \in \omega^{[n]}$ for all $n > n(x)$. Finally, let $\Omega^{[n]} = \omega^{[n]} \times (-h/2, h/2)$ and $\Omega^{[0]} = \omega^{[0]} \times (-h/2, h/2)$.

Assume now that the unit load is imposed, i.e., $f = p/D = 1$ (see § 2). Then for fixed thickness h there exist the unique solutions $\underline{u}_0^{[n]}$, $(w_R^{[n]}, \underline{\varrho}_R^{[n]})$, and $w_K^{[n]}$, $n = 0, 1, 2, \cdots$, corresponding, respectively, to the three-dimensional, Reissner-Mindlin, and Kirchhoff formulations of the plate-bending problem with either hard or soft

simple support. In § 4.1 we will show that $w_K^{[n]} \to w_K^{[\infty]} \neq w_K^{[0]}$ and give explicit expressions for $w_K^{[\infty]}$ and $w_K^{[0]}$. This is the plate paradox in the Kirchhoff model pointed out in [7] and [3]. In § 4.2 we will show that this paradox also occurs in the Reissner-Mindlin model and in the three-dimensional formulation in case of *hard simple support.* Finally, in § 4.3 we show that the paradox *does not occur* in the Reissner-Mindlin and three-dimensional formulations where soft simple support is imposed. This has been briefly noted in [3].

The results clearly show that seemingly minor changes in the boundary conditions can lead to a significant change of the solution on $\Omega^{[n]}$, respectively, $\omega^{[n]}$, when n is large. In fact, we will see that there can be significant changes already when $n = 1$.

The main question we will address below in this section is whether, as $n \to \infty$,

$$\underline{u}^{[n]} \to \underline{u}^{[0]} \quad \text{for the three-dimensional formulation,}$$

$$(w_R^{[n]}, \underline{\theta}_R^{[n]}) \to (w_R^{[0]}, \underline{\theta}_R^{[0]}) \quad \text{for the Reissner-Mindlin model,}$$

$$w_k^{[n]} \to w_k^{[0]} \quad \text{for the Kirchhoff model.}$$

4.1. The plate paradox for the Kirchhoff model. We have shown in Lemma 3.6 that for $n = 1, 2, \cdots$, $w_k^{[n]} = \rho^{[n]}$, and $-\Delta w_k^{[n]} = \psi^{[n]}$, where $\rho^{[n]}$ and $\psi^{[n]}$ satisfy (3.10a, b).

THEOREM 4.1. *Let* $\rho^{[\infty]}, \psi^{[\infty]} \in H_0^1(\omega^{[0]})$ *be such that*

$$(\underline{\nabla}\rho^{[\infty]}, \underline{\nabla}\xi) = (\psi^{[\infty]}, \xi), \qquad \xi \in H_0^1(\omega^{[0]}), \tag{4.1a}$$

$$(\underline{\nabla}\psi^{[\infty]}, \underline{\nabla}\xi) = \langle f, \xi \rangle, \qquad \xi \in H_0^1(\omega^{[0]}) \tag{4.1b}$$

with $f = 1$. *Then as* $n \to \infty$

$$\|\psi^{[n]} - \psi^{[\infty]}\|_{H^1(\omega^{[0]})} \to 0, \qquad \|\varphi^{[n]} - \rho^{[\infty]}\|_{H^1(\omega^{[0]})} \to 0.$$

Here we understand that $\psi^{[n]}$ *and* $\varphi^{[n]}$ *extend by zero from* $\omega^{[n]}$ *to* $\omega^{[0]}$.

Proof. Let P_n denote the orthogonal projection of $H_0^1(\omega^{[0]})$ onto the subspace $H_0^{1,n}(\omega^{[0]})$ defined by

$$H^{1,n}(\omega^{[0]}) = \{u \in H_0^1(\omega^{[0]}) \colon u = 0 \text{ on } \omega^{[0]} - \omega^{[n]}\}$$

and let $\hat{\psi}^{[n]}$ and $\hat{\rho}^{[n]}$ denote the extension of $\psi^{[n]}$ and $\rho^{[n]}$, respectively, by zero onto $\omega^{[0]}$. Then $\hat{\psi}^{[n]} = P_n \psi^{[\infty]}$ by (4.1b). From Theorem C.1 in Appendix C, it then follows immediately that $\hat{\psi}^{[n]} \to \psi^{[\infty]}$ in $H_0^1(\omega^{[0]})$. From (4.1a) we then see that $\rho^{[n]} - P_n \rho^{[\infty]} \to 0$ in $H_0^1(\omega^{[0]})$, and therefore, by the same argument, that $\hat{\rho}^{[n]} \to \rho^{[\infty]}$ in $H_0^1(\omega^{[0]})$. □

Let us now characterize $\hat{\rho}^{[0]} = w_K^{[0]}$ and $\rho^{[\infty]} = w_K^{[\infty]}$ more explicitly. To this end, note first that $\rho^{[\infty]}$ is the solution to the problem

$$\Delta\Delta\rho^{[\infty]} = 1 \qquad \text{on } \omega^{[0]}, \tag{4.2a}$$

$$\rho^{[\infty]} = \Delta\rho^{[\infty]} = 0 \quad \text{on } \partial\omega^{[0]}. \tag{4.2b}$$

On the other hand, it is easy to see that $\rho^{[0]}$ is the solution of the problem

$$\Delta\Delta\rho^{[0]} = 1 \quad \text{on } \omega^{[0]}, \tag{4.3a}$$

$$\rho^{[0]} = 0 \qquad \text{on } \partial\omega^{[0]}, \tag{4.3b}$$

$$\nu\Delta\rho^{[0]} + (1-\nu)\frac{\partial^2 \rho^{[0]}}{\partial n^2} = 0. \tag{4.3c}$$

Here (4.3c) is the standard boundary condition for the simply supported circular plate (see, e.g., [24, p. 554]). Solutions (4.2) and (4.3) show that

$$\rho^{[\infty]} = C_1^{[\infty]} + C_2^{[\infty]} r^2 + C_3^{[\infty]} r^4, \qquad \rho^{[0]} = C_1^{[0]} + C_2^{[0]} r^2 + C_3^{[0]} r^4,$$

where $r^2 = x_1^2 + x_2^2$,

$$C_3^{[\infty]} = C_3^{[0]} = \frac{1}{64},$$

and C_1, C_2 are determined from the boundary conditions. By simple computation we get

$$\rho^{[0]}(0,0) = w_K^{[0]}(0,0) = \frac{1}{64}\frac{5+\nu}{1+\nu}, \tag{4.4a}$$

$$\rho^{[\infty]}(0,0) = w_K^{[\infty]}(0,0) = \frac{3}{64}, \tag{4.4b}$$

and hence for $\nu = 0.3$ we have

$$\frac{w_K^{[0]}(0,0)}{w_K^{[\infty]}(0,0)} = 1.36,$$

i.e., the gap between $w_K^{[0]}$ and $w_K^{[\infty]}$ is 36 percent at the origin. Analogously, for $\nu = 0.3$,

$$\frac{\|w_K^{[0]} - w_K^{[\infty]}\|_{0,\omega^{[0]}}}{\|w_K^{[\infty]}\|_{0,\omega^{[0]}}} = 0.287.$$

Remark 4.1. We have assumed that $\omega^{[n]}$ were regular polygons. As the proof shows, (4.5b) also holds when $\{\omega^{[n]}\}$ is an arbitrary sequence of *convex* polygons such that $\omega^{[n]} \to \omega^{[0]}$ in the sense described above.

It is essential, however, that $\omega^{[n]}$ are convex polygons. If we replace $\omega^{[n]}$ by $\hat{\omega}^{[n]}$, where $\hat{\omega}^{[n]}$ are nonconvex polygons as shown in Fig. 4.1, then [15] shows that $\hat{\omega}^{[\infty]}$

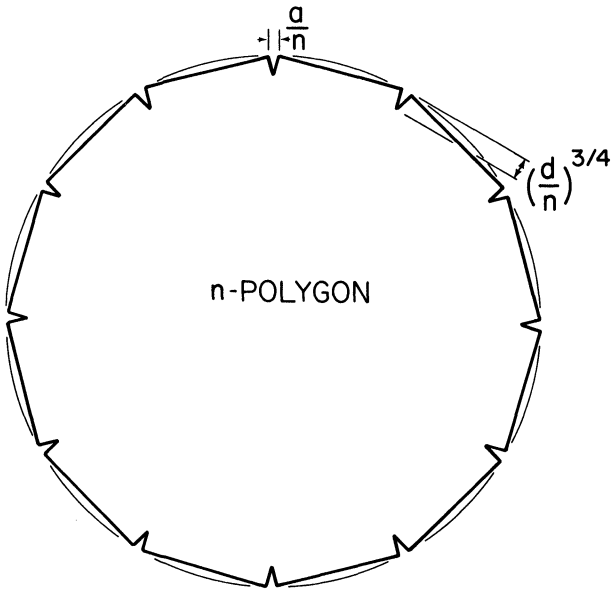


FIG. 4.1. *A nonconvex polygon* $\omega^{[n]}$.

satisfies

$$\Delta\Delta\hat{\omega}^{[\infty]} = 1 \qquad \text{in } \omega^{[0]},$$

$$\hat{\omega}^{[\infty]} = \frac{\partial\hat{\omega}^{[\infty]}}{\partial n} = 0 \qquad \text{in } \partial\omega^{[0]},$$

and hence

$$\hat{\omega}^{[\infty]}(0,0) = \frac{1}{64}. \tag{4.4c}$$

4.2. The plate paradox for the three-dimensional and Reissner–Mindlin models. We will analyze in detail the case of Reissner-Mindlin model only. The case of the three-dimensional formulation can be dealt with analogously.

THEOREM 4.2. *Let h be fixed and sufficiently small, and let $w_R^{[n]}$ be the Reissner-Mindlin solution on $\omega^{[n]}$ corresponding to unit load $f = 1$ on $\omega^{[n]}$ and hard simple support on $\partial\omega^{[n]}$, $n = 0, 1, 2, \cdots$. Then if $w_R^{[n]}$ is extended by zero onto $\omega^{[0]}$, we have*

$$\|w_R^{[n]} - \omega_R^{[0]}\|_{1,\omega^{[0]}} \geqq \alpha > 0,$$

$$\left| \int_{\omega^{[0]}} (w_R^{[n]} - w_R^{[0]})\, dx_1\, dx_2 \right| \geqq \alpha > 0$$

for all $n \geqq n_0$, n_0 large enough.

Proof. By Theorem 3.1 we have

$$|||w_R^{[n]} - w_K^{[n]}, \underline{\theta}_R^{[n]} - \underline{\theta}_K^{[n]}, \underline{m}_R^{[n]} - \underline{m}_K^{[n]}, \underline{\gamma}_R^{[n]} - \underline{\gamma}_K^{[n]}|||_R^2 \leqq h^2/\kappa \|f\|_{-1,\omega^{[n]}}^2,$$

$$|||w_R^{[0]} - w_K^{[0]}, \underline{\theta}_R^{[0]} - \underline{\theta}_K^{[0]}, \underline{m}_R^{[0]} - \underline{m}_K^{[0]}, \underline{\gamma}_R^{[0]} - \underline{\gamma}_K^{[0]}|||_R^2 \leqq Ch^2/\kappa \|f\|_{-1,\omega^{[0]}}^2.$$

Note that $\|f\|_{-1,\omega^{[n]}} \leqq C_0$ independently of n. Using Lemma 3.3 and Theorem B.3, we see that

$$[\|w_R^{[n]} - w_K^{[n]}\|_{1,\omega^{[n]}}^2 + \|\underline{\theta}_R^{[n]} - \underline{\theta}_K^{[n]}\|_{1,\omega^{[n]}}^2] \leqq Ch^2,$$

$$[\|w_R^{[0]} - w_K^{[0]}\|_{1,\omega^{[0]}}^2 + \|\underline{\theta}_R^{[0]} - \underline{\theta}_K^{[0]}\|_{1,\omega^{[0]}}^2] \leqq Ch^2,$$

where C is independent of n and h. On the other hand, we have by Theorem 4.1,

$$\|w_K^{[n]} - w_K^{[\infty]}\|_{1,\omega^{[0]}} \to 0 \quad \text{as } n \to \infty,$$

$$\|w_K^{[\infty]} - w_K^{[0]}\|_{1,\omega^{[0]}} > 0.$$

This shows that for sufficiently small h there is $\alpha > 0$ such that $\|w_R^{[0]} - w_R^{[n]}\|_{1,\omega^{[n]}} \geqq \alpha > 0$ for all $n > n_0$.

Realizing that (in our case for $f = 1$)

$$E_R^{[n]} = -\int_{\omega^{[n]}} w_R^{[n]}\, dx_1\, dx_2, \qquad E_R^{[0]} = -\int w_R^{[0]}\, dx_1\, dx_2,$$

$$E_K^{[n]} = -\int w_K^{[n]}\, dx_1\, dx_2, \qquad E_R^{[0]} = -\int w_K^{[0]}\, dx_1\, dx_2,$$

we also have

$$\left| \int (\omega_R^{[0]} - w_R^{[n]})\, dx_1\, dx_2 \right| \geqq \alpha > 0 \quad \text{as } n > n_0. \qquad \square$$

Using Theorem 3.2 and analogous arguments, we get Theorem 4.3.

THEOREM 4.3. *Let h be fixed and sufficiently small and let $\underline{u}_0^{[n]} = (u_{01}^{[n]}, u_{02}^{[n]}, u_{03}^{[n]})$ be the three-dimensional solution of the plate-bending problem on $\Omega^{[n]}$ corresponding to*

the load $p = D$ and hard simple support, $n = 0, 1, 2, \cdots$. Then if $u_{03}^{[n]}$ is extended by zero onto $\Omega^{[0]}$, we have

$$\|u_{03}^{[n]} - u_{03}^{[0]}\|_{1,\Omega^{[0]}} \geqq \alpha > 0,$$

$$\left| \int_{\omega^{[0]}} \left(u_{03}^{[n]}\left(x_1, x_2, \frac{h}{2}\right) + u_{03}^{[0]}\left(x_1, x_2, \frac{h}{2}\right)\right) dx_1\, dx_2 \right| \geqq \alpha > 0$$

for all $n \geqq n_0$, n_0 sufficiently large.

Theorems 4.2 and 4.3 show that hard simple support leads not only to the paradox in the Kirchhoff model but also to those in the three-dimensional formulation and the Reissner–Mindlin model. (In § 4.3 we will show that the paradox occurs neither in the three-dimensional formulation nor in the Reissner–Mindlin model when the simple soft support is imposed.)

The proof employs the fact that the Kirchhoff model approximates very well the Reissner–Mindlin and three-dimensional formulations for the *hard* support. This shows that the circular plate and polygonal plate solutions are far apart in the entire region and not only in the area close to the boundary, where boundary layer effects occur.

The results above show that plausibly unimportant changes in the boundary conditions could lead to significant changes in the solution through the entire region even if the three-dimensional linear elasticity model is used. We expect that the paradox will also occur in nonlinear formulations. For engineering implications of effects of this type we refer to [6].

4.3. The "nonparadox" in case of soft simple support. In this section we will prove that, in contrast to hard simple support, the solution on $\omega^{[n]}$ converges to the solution on $\omega^{[0]}$ for both the Reissner–Mindlin and the three-dimensional plate model. This is in obvious contrast to hard simple support. We will elaborate in detail on the case of the Reissner–Mindlin model. The analysis of the three-dimensional model is analogous.

Let us denote

$$\mathscr{D}_n = \omega^{[n+1]} - \omega^{[n]}, \qquad n = 1, 2, \cdots,$$

$$\mathscr{D}_0 = \omega^{[1]},$$

$$\mathscr{D}_n^0 = \omega^{[0]} - \omega^{[n]}, \qquad n = 1, 2, \cdots$$

(see Fig. 4.2).

Let $L = (L_2(\omega^{[0]})^3$, $\underline{u} = (w, \underline{\theta}) \in L$ and

$$\mathscr{S}_0 = \{\underline{u} \in L\colon w \in H_0^1(\omega^{[0]}),\ \underline{\theta} \in (H^1(\omega^{[0]}))^2\},$$

$$\mathscr{S}_n = \{u \in L\colon w \in H_0^1(\omega^{[0]}),\ w = 0 \text{ on } \mathscr{D}_n^0,\ \underline{\theta} \in (H^1(\omega^{[0]})^2)\},$$

$$\mathscr{T}_n = \{u \in L\colon w \in H_0^1(\omega^{[0]}),\ \underline{\theta} \in (H^1(\omega^{[n]}))^2,\ \underline{\theta} \in (H^1(\mathscr{D}_m))^2,\ m = n, n+1, \cdots\},$$

$$\mathscr{L}_{n,m} = \{u \in L\colon w \in H_0^1(\omega^{[0]}),\ w = 0 \text{ on } \mathscr{D}_n^0,$$

$$\underline{\theta} \in (H^1(\omega^{[m]}))^2,\ \underline{\theta} \in (H^1(\mathscr{D}_j))^2,\ j = m, m+1, \cdots\}.$$

We have $\mathscr{S}_n \subset \mathscr{S}_0$, $\mathscr{S}_0 \subset \mathscr{T}_n$, and

$$\mathscr{L}_{n,m} \supset \mathscr{S}_n, \qquad \mathscr{L}_{n,m} \subset \mathscr{T}_m.$$

All the spaces are embedded in $\mathscr{T}_1$. Furthermore, let

$$Z_n = \{u \in L\colon w \in H^1(\omega^{[n]}),\ \underline{\theta} \in (H^1(\omega^{[n]})^2\},$$

$$\mathring{Z}_n = \{u \in Z_n\colon w \in H_0^1(\omega^{[n]})\},$$

$$\mathscr{A}_R^0(w, \underline{\theta}; z, \underline{\varphi}) = \sum_{i=0}^{\infty} \mathscr{A}_R^{\mathscr{D}_i}(\underline{u}, \underline{v}),$$

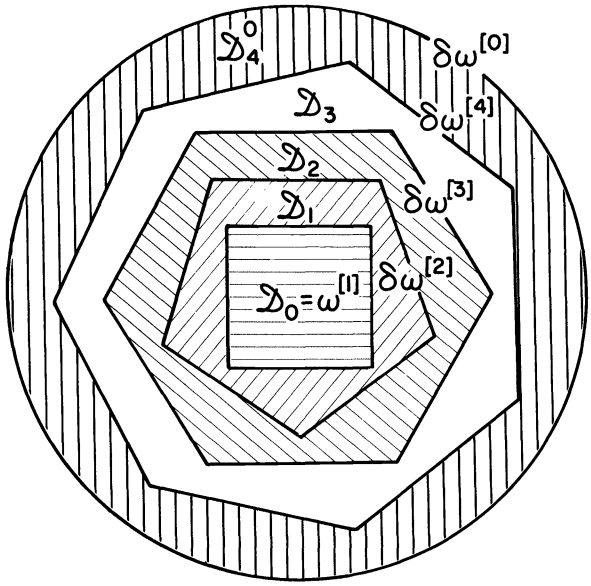


FIG. 4.2. *The configuration of the domains* $\mathcal{D}_n$, $\mathcal{D}_0$, $\mathcal{D}'_n$.

where $\mathcal{A}_R^{\omega}$ is given in § 3.2 for the region ω and $\mathcal{A}_R^{\mathcal{D}^i}$ has the same form but is integrated only over $\mathcal{D}^i$. Analogously we define $\mathcal{A}_R^{\omega^{[n]}}$, etc. Finally, we supply $\mathcal{T}_1$ with the norm

$$\|\underline{u}\|^2 = \sum_{i=0}^{\infty} \mathcal{A}_R^{\mathcal{D}^i}(\underline{u}, \underline{u}).$$

To see that $\|\cdot\|$ is indeed a norm, assume that $\underline{u} = (w, \underline{\theta}) \in \mathcal{T}_1$ and $\|\underline{u}\| = 0$. Then, since the first term in the expression for $\mathcal{A}_R^{\mathcal{D}^i}$ is the same as in the case of plane elasticity (where θ_1, θ_2 play the role of the displacements), we have on $\mathcal{D}_j$, $\theta_1 = a_j + c_j x_2$, $\theta_2 = b_j - c_j x_1$, and because $\|\underline{\theta} - \nabla w\|_{0,\mathcal{D}_j} = 0$ we get $c_j = 0$. Hence $w = d_j + a_j x_1 + b_j x_2$ on $\mathcal{D}_j$, and so, because $w \in H_0^1(\omega^{[0]})$ we get $w = 0$ and $a_j = b_j = 0$, $j = 0, 1, 2, \cdots$ (see also Appendix B). Hence $\underline{u} = 0$ and accordingly, $\|\cdot\|$ is a norm on $\mathcal{T}_1$.

For $\underline{u} \in Z_n$ let $\|\underline{u}\|^2_{R,\omega^{[n]}} = \mathcal{A}_R^{\omega^{[n]}}(\underline{u}, \underline{u})$. Then by Theorem B.1 in Appendix B,

$$\inf_{abc} \|\theta_1 - (a + cx_2), \theta_2 - (b - cx_1)\|_{1,\omega^{[n]}} \leqq C_n \|u\|_{R,\omega^{[n]}}, \tag{4.5a}$$

$$\inf_{abcd} \|w - (d + ax_1 + bx_2 + cx_1 x_2)\|_{1,\omega^{[n]}} \leqq C_n \|u\|_{R,\omega^{[n]}}. \tag{4.5b}$$

Here C_n depends in general on $\omega^{[n]}$.

Assume now that for an $n_0 > 0$

(4.6a) f has compact support in $\omega^{[n_0]}$,

$$\int_{\omega^{[n_0]}} f\,dx_1\,dx_2 = \int_{\omega^{[n_0]}} fx_1\,dx_1\,dx_2 = \int_{\omega^{[n_0]}} fx_2\,dx_1\,dx_2 = \int_{\omega^{[n_0]}} fx_1 x_2\,dx_1\,dx_2 = 0 \tag{4.6b}$$

and that $n > n_0$, $m > n_0$. Then for $\underline{u} \in \mathcal{T}_n$, $n \geqq n_0$,

$$\left| \int_{\omega^{[0]}} fw\,dx_1\,dx_2 \right| = \left| \int_{\omega^{[n_0]}} fw\,dx_1\,dx_2 \right| \leqq C_{n_0} \|u\|.$$

Hence for $n, m \geqq n_0$ there exist unique

$$\underline{u}(\mathcal{S}_n) \in \mathcal{S}_n, \quad \underline{u}(\mathcal{T}_n) \in \mathcal{T}_n, \quad \underline{u}(\mathcal{L}_{n,m}) \in \mathcal{L}_{n,m}, \quad \underline{u}(\mathring{\mathcal{Z}}_n) \in \mathring{\mathcal{Z}}_m$$

such that

$$t_R^0(\underline{u}(\mathcal{S}_n), \underline{v}) = \int_{\omega^{[0]}} fz \, dx_1 \, dx_2 \quad \forall v \in (z, \underline{\varphi}) \in \mathcal{S}_n$$

and analogously for $u(\mathcal{T}_n)$, $u(\mathcal{L}_{n,m})$, $u(\mathring{\mathcal{Z}}_n)$. Obviously $\underline{u}(\mathcal{S}_0) = \underline{u}_R^{[0]}$ and $\underline{u}(\mathring{\mathcal{Z}}_n) = \underline{u}_R^{[n]}$, and $\underline{u}(\mathcal{L}_{n,m}) = \underline{u}(\mathring{\mathcal{Z}}_n)$ on $\omega^{[n]}$ and is zero on $\mathcal{D}_n^0$.

Using Theorem C.1 we get

(4.7a) $$\underline{u}_R^{[0]} = \underline{u}(\mathcal{S}_0) = \underline{u}(\mathcal{S}_n) + \underline{\rho}(\mathcal{S}_0, \mathcal{S}_n),$$

(4.7b) $$\|\underline{u}(\mathcal{S}_0)\|^2 = \|\underline{u}(\mathcal{S}_n)\|^2 + \|\underline{\rho}(\mathcal{S}_0, \mathcal{S}_n)\|^2,$$

(4.7c) $$\|\underline{\rho}(\mathcal{S}_0, \mathcal{S}_n)\| \to 0 \quad \text{as } n \to \infty;$$

(4.8a) $$\underline{u}(\mathcal{T}_n) = \underline{u}(\mathcal{S}_0) + \underline{\rho}(\mathcal{T}_n, \mathcal{S}_0),$$

(4.8b) $$\|\underline{u}(\mathcal{T}_n)\|^2 = \|\underline{u}(\mathcal{S}_0)\|^2 + \|\underline{\rho}(\mathcal{T}_n, \mathcal{S}_0)\|^2,$$

(4.8c) $$\|\underline{\rho}(\mathcal{T}_n, \mathcal{S}_0)\| \to 0 \quad \text{as } n \to \infty;$$

(4.9a) $$\underline{u}(\mathcal{L}_{n,m}) = \underline{u}(\mathcal{S}_n) + \underline{\rho}(\mathcal{L}_{n,m}, \mathcal{S}_n),$$

(4.9b) $$\|\underline{u}(\mathcal{L}_{n,m})\|^2 = \|\underline{u}(\mathcal{S}_n)\|^2 + \|\underline{\rho}(\mathcal{L}_{n,m}, \mathcal{S}_n)\|^2,$$

(4.9c) $$\|\underline{\rho}(\mathcal{L}_{n,m}, \mathcal{S}_n)\| \to 0 \quad \text{as } m \to \infty;$$

(4.10a) $$\underline{u}(\mathcal{T}_m) = \underline{u}(\mathcal{L}_{n,m}) + \underline{\rho}(\mathcal{T}_m, \mathcal{L}_{n,m}),$$

(4.10b) $$\|\underline{u}(\mathcal{T}_m)\|^2 = \|\underline{u}(\mathcal{L}_{n,m})\|^2 + \|\underline{\rho}(\mathcal{T}_m, \mathcal{L}_{n,m})\|^2,$$

(4.10c) $$\|\underline{\rho}(\mathcal{T}_m, \mathcal{L}_{n,m})\| \to 0 \quad \text{as } n \to \infty.$$

Now let $\varepsilon > 0$ and $n > \max(n(\varepsilon), n_0)$. Then we have

$$\|\underline{\rho}(\mathcal{S}_0, \mathcal{S}_n)\|^2 < \varepsilon, \qquad \|\underline{\rho}(\mathcal{T}_n, \mathcal{S}_0)\|^2 < \varepsilon.$$

Using (4.7)-(4.10) we get

$$\begin{aligned}\|\underline{u}(\mathcal{S}_0)\|^2 + \|\underline{\rho}(\mathcal{T}_m, \mathcal{S}_0)\|^2 &= \|\underline{u}(\mathcal{T}_m)\|^2 = \|\underline{u}(\mathcal{L}_{n,m})\|^2 + \|\underline{\rho}(\mathcal{T}_m, \mathcal{L}_{n,m})\|^2 \\ &= \|\underline{u}(\mathcal{S}_n)\|^2 + \|\underline{\rho}(\mathcal{L}_{n,m}, \mathcal{S}_n)\|^2 + \|\underline{\rho}(\mathcal{T}_n, \mathcal{L}_{n-n})\|^2 \\ &= \|\underline{u}(\mathcal{S}_n)\|^2 - \|\underline{\rho}(\mathcal{S}_0, \mathcal{S}_n)\|^2 + \|\underline{\rho}(\mathcal{L}_{n,m}, \mathcal{S}_n)\|^2 \\ &\quad + \|\underline{\rho}(\mathcal{T}_m, \mathcal{L}_{n,m})\|^2\end{aligned}$$

and hence for $n, m \geqq \max(n(\varepsilon), n_0)$

$$\|\underline{\rho}(\mathcal{T}_m, \mathcal{S}_0)\|^2 + \|\underline{\rho}(\mathcal{S}_0, \mathcal{S}_n)\|^2 = \|\underline{\rho}(\mathcal{L}_{n,m}, \mathcal{S}_n)\|^2 + \|\underline{\rho}(\mathcal{T}_m, \mathcal{L}_{n,m})\|^2 \leqq 2\varepsilon,$$

which yields

$$\|\underline{\rho}(\mathcal{L}_{n,n}, \mathcal{S}_n)\|^2 \leqq 2\varepsilon.$$

Therefore

$$\underline{u}(\mathcal{S}_0) - \underline{u}(\mathcal{L}_{n,n}) = \underline{u}(\mathcal{S}_0) - \underline{u}(\mathcal{S}_n) + \underline{u}(\mathcal{S}_n) - \underline{u}(\mathcal{L}_{n,n}) = \underline{\rho}(\mathcal{S}_0, \mathcal{S}_n) - \underline{\rho}(\mathcal{L}_{n,n}, \mathcal{S}_n)$$

and hence

$$\|\underline{u}(\mathcal{S}_0) - \underline{u}(\mathcal{L}_{n,n})\| \leqq \varepsilon^{1/2} + \sqrt{2\varepsilon} \leqq C\varepsilon^{1/2}.$$

Because, as above, $\underline{u}(\mathcal{L}_{n,n}) = \underline{u}_R^{[n]}$ on $\omega^{[n]}$, and zero on $\mathcal{D}_n^0$, $\underline{u}_R^{[n]} \to \underline{u}_R^{[0]}$ in the space $\mathcal{T}_1$ or in any $\mathcal{T}_m$ for m fixed.

Remark 4.2. Note that until now we have not used Theorem B.3 (Appendix B), but only Theorem B.1.

So far we have assumed that f satisfies (4.6). Let us now study the general case. Assume that $f \in L_2(\omega^{[0]})$.

Let us first note that if $\underline{u} = (w, \theta) \in \mathscr{L}_{n,n}$, then $w \in H_0^1(\omega^{[0]})$ and

$$\|w\|_{1,\omega^{[n]}} = \|w\|_{1,\omega^{[n]}} \leqq C\|\underline{u}\| \tag{4.11}$$

with C independent of n because of Theorem B.3.

For $0 < \Delta < \frac{1}{2}$ we denote

$$R_\Delta = \{(x_1, x_2): x_1^2 + x_2^2 > 1 - \Delta\},$$
$$\partial R_\Delta = \{(x_1, x_2): x_1^2 + x_2^2 = 1 - \Delta\}.$$

Then

$$\|w\|_{0,R_\Delta} \leqq C\Delta \|w\|_{1,\omega^{[0]}} \leqq C\Delta\|u\|,$$
$$\|w\|_{0,\partial R_\Delta} \leqq C\Delta^{1/2}\|w\|_{1,\omega^{[0]}} \leqq C\Delta^{1/2}\|u\|.$$

Now let

$$f_\Delta = \begin{cases} f & \text{on } R_\Delta, \\ 0 & \text{on } \omega^{[0]} - R_\Delta, \end{cases}$$
$$g_\Delta = (a + bx_1 + cx_2 + dx_1x_2)\mathscr{O}_\Delta,$$

where $\mathscr{O}_\Delta$ is the Dirac function concentrated on ∂R_Δ and a, b, c, d are such that $f_\Delta + g_\Delta$ satisfies (4.6).

For $n > n_{1,\Delta}$ such that $\bar{\bar{R}}_\Delta \subset \omega^{[n_{1,\Delta}]}$, let $\underline{u}_\Delta(\mathscr{L}_{n,n})$ and $\underline{u}_\Delta(\mathscr{S}_0)$ be the solutions when instead of f the function f_Δ is used. Then we get

$$\|\underline{u}_\Delta(\mathscr{L}_{n,n}) - \underline{u}(\mathscr{L}_{n,n})\| \leqq C\Delta^{1/2},$$
$$\|\underline{u}_\Delta(\mathscr{S}_0) - \underline{u}(\mathscr{S}_0)\| \leqq C\Delta^{1/2},$$

where C is independent of n and Δ but, in general, depends on f. Hence we can select Δ so that $C\Delta^{1/2} < \varepsilon$. Furthermore, we have shown

$$\|\underline{u}_\Delta(\mathscr{L}_{n,n}) - \underline{u}_\Delta(\mathscr{S}_0)\| < \varepsilon$$

for all $n \geqq n_1(\varepsilon)$ and therefore

$$\|\underline{u}(\mathscr{L}_{n,n}) - \underline{u}(\mathscr{S}_0)\| < 3\varepsilon$$

for all $n \geqq n_1(\varepsilon)$. Since $\underline{u}(\mathscr{S}_0) = \underline{u}_R^{[0]}$ and $\underline{u}(\mathscr{L}_{n,n}) = \underline{u}_R^{[n]}$, we get

$$\|\underline{u}_R^{[0]} - \underline{u}_R^{[n]}\| \to 0 \quad \text{as } n \to \infty.$$

Here $\underline{u}_R^{[n]} = (w_R^{[n]}, \underline{\theta}_R^{[n]})$ is understood to be extended by zero on $\mathscr{D}_n^0$ and $\|\cdot\|$ is the norm in $\mathscr{T}_1$ (note that $w_R^{[n]} \in H_0^1(\omega^{[0]})$, but $\underline{\theta}_R^{[n]} \notin H^1(\omega^{[0]})$ although $\theta_R^{[n]} \subset H^1(\omega^{[n]})$). Because the functions in $H^1(\omega^{[0]})$ with compact support are dense in $H_0^1(\omega^0)$, there is $\bar{w}^{[n]} \in H_0^1(\omega^{[n]})$ such that $\|w^{[0]} - \bar{w}^{[n]}\| \leqq \varepsilon$ for all $n \geqq n_2(\varepsilon)$. Hence with $\bar{\underline{u}}^{[n]} = (\bar{w}^{[n]}, \underline{\theta}_R^{[n]})$ we get

$$\|\bar{\underline{u}}^{[n]} - \underline{u}_R^{[0]}\| \to 0 \quad \text{as } n \to \infty.$$

Hence, using Theorem B.2 (Appendix B), we have

$$\|w_R^{[n]} - w_R^{[0]}\|^2_{H^1(\omega^{[n]})} + \|\underline{\theta}^{[n]} - \underline{\theta}_R^{[0]}\|^2_{H^1(\omega^{[n]})} \to 0 \quad \text{as } n \to \infty.$$

In summary, we have proved Theorem 4.4 below.

THEOREM 4.4. *Let* $f \in L_2(\omega^{[0]})$ *and let* $u_R^{[n]} = (w_R^{[n]}, \underline{\theta}_R^{[n]})$, *respectively,* $u_R^{[0]} = (w_R^{[0]}, \underline{\theta}_R^{[0]})$, *be the Reissner-Mindlin solution on* $\omega^{[n]}$, *respectively,* $\omega^{[0]}$, *for soft simple support and fixed h. Then*

$$\|w_R^{[n]} - w_R^{[0]}\|_{H^1(\omega^{[n]})} + \|\underline{\theta}_R^{[n]} - \underline{\theta}_R^{[0]}\|_{H^1(\omega^{[n]})} \to 0 \quad \text{as } n \to \infty.$$

We see that in contrast to the hard support there is no plate paradox when the soft support is imposed. Hence soft simple support is physically more natural than hard simple support.

Remark 4.3. In Theorem 4.2 we assumed that $f \in L_2(\omega^{[0]})$ while the solutions $\underline{u}_R^{[0]}$ and $\underline{u}_R^{[n]}$ were defined for any $f \in H^{-1}(\omega^{[0]})$, respectively, $f \in H^{-1}(\omega^{[n]})$. If f has compact support, then Theorem 4.4 also holds for $f \in H^{-1}(\omega^{[0]})$. We can weaken the assumptions on f in Theorem 4.4, e.g., so that $f \in H^{\alpha}(\omega^{[0]})$, $\alpha > -\frac{1}{2}$, but the proof will not hold for $f \in H^{-1}(\omega^{[0]})$.

Remark 4.4. We have assumed that $\omega^{[n]}$ is the sequence of regular polygons. This assumption was used only when we were using Theorem B.3. Hence Theorem 4.4 holds for any regular family of domains (see Appendix B). If f satisfies (4.6), then there is no need for regularity (see Remark 4.2) of the family of domains under consideration and Theorem 4.4 holds in the full generality.

Remark 4.5. We have assumed in Theorem 4.5 that $h > 0$ is fixed (i.e., independent of n). We could also consider a two-parameter family of problems where both n and h vary. Then, for n fixed and $h \to 0$, $\underline{u}_R^{[n]} \to \underline{u}_K^{[n]}$ (and hence for $h \to 0$ the difference between soft and hard support disappears). Hence, combining the results of this section with § 4.2, we see that

$$\lim_{n\to\infty} \lim_{h\to 0} \underline{u}_R^{(n)} \neq \lim_{h\to 0} \lim_{n\to\infty} \underline{u}_R^{(n)}.$$

In a way analogous to the proof of Theorem 4.4, we can prove Theorem 4.5.

THEOREM 4.5. *Let h be fixed and* $\underline{u}^{[0]}$, *respectively* $\underline{u}^{[n]}$, *be the solution of the three-dimensional plate problem on* $\Omega^{[0]}$, *respectively,* $\Omega^{[n]}$, *with soft simple support. Assume that the load* $p \in L_2(\omega^{[0]})$. *Then*

$$\|\underline{u}^{[0]} - \underline{u}^{[n]}\|_{1,\Omega^{[n]}} \to 0$$

as $n \to \infty$.

Remark 4.6. Remarks 4.3–4.5 are also valid for the three-dimensional plate model.

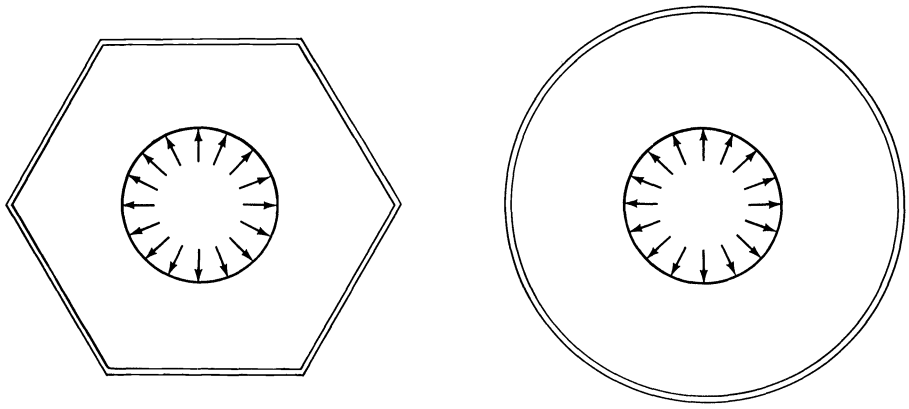
4.4. Some additional considerations. As we have seen, the Kirchhoff model (biharmonic equation) leads to paradoxical behavior for hard simple support. The same mathematical formulation also describes other problems and hence leads to the same paradoxical behavior.

As an example, we mention the problem of a reinforced tube shown in Fig. 4.3a, b. The reinforcement is attached by an unextendable tape to the exterior surface. Here we have the paradox that the stress caused by hydrostatic pressure is different for the polygonal and circular outer surfaces.

Analogous examples can very likely be found in fields other than elasticity where the problem reduces to the biharmonic (or polyharmonic) equation.

We have shown the paradoxical behavior for $n \to \infty$ and h relatively large compared with $1/n$ (see Remark 4.5). Hence the question arises of how large will be the difference between hard and soft support in three-dimensional formulation for n fixed and $h \to 0$. To this end we consider a square plate with sidelength equal to 1. In Table 4.1 we give the values of

$$\left[\frac{|E_{\text{SOFT}} - E_{\text{HARD}}|}{|E_{\text{SOFT}}|}\right]^{1/2} = \eta(h), \qquad \left[\frac{|E_{\text{HARD}} - E_K|}{|E_{\text{HARD}}|}\right]^{1/2} = \xi(h).$$



(a) (b)

FIG. 4.3. *Reinforced polygonal and circular tubes.*

TABLE 4.1

Percent	$h = 0.1$	$h = 0.01$
η	34.68	11.69
ξ	20.21	2.03

Here by E_{SOFT} and E_{HARD} we denote the (three-dimensional) plate energy for soft and hard support, and by E_K, the plate energy of the Kirchhoff model for the Poisson ratio $\nu = 0$ (see also [7]).

Appendix A. Well-posedness of variational problems (3.4), (3.7), and (3.8). We use the following basic theorem (see [18]).

THEOREM A.1. *Let H be a Hilbert space and $\mathcal{B}$ be a bilinear form on $H \times H$ that satisfies*

$$\text{(A0)} \qquad \mathcal{B}(u, v) = \mathcal{B}(v, u), \qquad u, v \in H,$$

$$\text{(A1)} \qquad |\mathcal{B}(u, v)| \leqq C\|u\|_H \|v\|_H, \qquad u, v \in H,$$

$$\text{(A2)} \qquad \sup_{\substack{v \in H \\ \|v\|_H = 1}} \mathcal{B}(u, v) \geqq c\|u\|_H \quad \forall u \in H,$$

where C and c are positive constants. Then if F is any bounded linear functional on H, there is a unique $u \in H$ satisfying

$$\text{(A3)} \qquad \mathcal{B}(u, v) = F(v), \qquad v \in H.$$

In applying Theorem A.1 to problems (3.4), (3.7), and (3.8), we choose the following notation.

(a) The three-dimensional model (3.4):

$$H = U \times \mathcal{H},$$

$$\mathcal{B}(\underline{u}, \underline{\sigma}; \underline{v}, \underline{\tau}) = (\underline{\sigma} S^{-1} \underline{\tau})_{\mathcal{H}} - (\underline{\varepsilon}(\underline{u}), \underline{\tau})_{\mathcal{H}} - (\underline{\sigma}, \underline{\varepsilon}(\underline{v}))_{\mathcal{H}},$$

$$F(\underline{v}, \underline{\tau}) = -Q(\underline{v}).$$

(b) The Reissner-Mindlin model (3.7):

$$H = H_0^1(\omega) \times V \times \mathcal{K} \times [L_2(\omega)]^2,$$

$$\mathcal{B}(w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}; z, \underline{\varphi}, \underline{\underline{k}}, \underline{\zeta}) = (\underline{\underline{m}}, T^{-1}\underline{\underline{k}})_{\mathcal{K}} - (\underline{\underline{\varepsilon}}(\underline{\theta}), \underline{\underline{k}})_{\mathcal{K}} - (\underline{\underline{\varepsilon}}(\underline{\varphi}), \underline{\underline{m}})_{\mathcal{K}} - (\underline{\theta} - \underline{\nabla} w, \underline{\zeta}) - (\underline{\varphi} - \underline{\nabla} z, \underline{\gamma}) + (h^2/\kappa)(\underline{\gamma}, \underline{\zeta}),$$

$$F(z, \underline{\varphi}, \underline{\underline{k}}, \underline{\zeta}) = -\langle f, z\rangle.$$

(c) The Kirchhoff model (3.8):

$$H = W \times V \times \mathcal{K} \times V',$$

$$\mathcal{B}(w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}; z, \underline{\varphi}, \underline{\underline{k}}, \underline{\zeta}) = (\underline{\underline{m}}, T^{-1}\underline{\underline{k}})_{\mathcal{K}} - (\underline{\underline{\varepsilon}}(\underline{\theta}), \underline{\underline{k}})_{\mathcal{K}} - (\underline{\underline{\varepsilon}}(\underline{\varphi}), \underline{\underline{m}})_{\mathcal{K}} - \langle \underline{\theta} - \underline{\nabla} w, \underline{\zeta}\rangle - \langle \underline{\varphi} - \underline{\nabla} z, \underline{\gamma}\rangle,$$

$$F(z, \underline{\varphi}, \underline{\underline{k}}, \underline{\zeta}) = -\langle f, z\rangle.$$

Then in each case, $\mathcal{B}$ is symmetric, F is a bounded linear functional on H, and the variational problem takes the general form (A3). Thus it suffices to show that (A1) and (A2) hold.

THEOREM A.2. *Assume that ω is a bounded Lipschitz domain and that the parameters ν, h, and κ satisfy*

$$0 \leqq \nu < \tfrac{1}{2}, \quad \bar{h} \leqq h \leqq \bar{h}^{-1}, \quad \bar{\kappa} \leqq \kappa \leqq \bar{\kappa}^{-1},$$

where $\bar{h} > 0$ and $\bar{\kappa} > 0$ are given. Then in each of the three cases above there are constants $C = C(\bar{h}, \bar{\kappa})$ and $c = c(\omega, \bar{h}, \bar{\kappa})$ such that (A1) *and* (A2) *hold.*

Proof. In view of (3.2) and (3.6) the mappings $S^{-1}: \mathcal{H} \to \mathcal{H}$ and $T^{-1}: \mathcal{K} \to \mathcal{K}$ are uniformly bounded in the assumed range of ν. It then follows easily that the assertion concerning (A1) holds, so let us concentrate on showing that (A2) is true.

(a) The three-dimensional model. Let $(\underline{u}, \underline{\underline{\sigma}}) \in U \times \mathcal{H}$ be given and let

$$(\underline{\underline{\sigma}}_0)_{ij} = \tfrac{1}{3} \operatorname{tr}(\underline{\underline{\sigma}}) \delta_{ij}, \qquad i, j = 1, 2, 3.$$

Then $\|\underline{\underline{\sigma}}\|_{\mathcal{H}}^2 = \|\underline{\underline{\sigma}} - \underline{\underline{\sigma}}_0\|_{\mathcal{H}}^2 + \frac{1}{3}\|\operatorname{tr}(\underline{\underline{\sigma}})\|_{0,\Omega}^2$ and it follows from (3.2) that

$$(\underline{\underline{\sigma}}, S^{-1}\underline{\underline{\sigma}})_{\mathcal{H}} = \frac{D}{E}\{(1+\nu)\|\underline{\underline{\sigma}} - \underline{\underline{\sigma}}_0\|_{\mathcal{H}}^2 + (1-2\nu)\|\underline{\underline{\sigma}}_0\|_{\mathcal{H}}^2\}$$

(A4)

$$\geqq \frac{h^3}{12}\|\underline{\underline{\sigma}} - \underline{\underline{\sigma}}_0\|_{\mathcal{H}}^2 \qquad (0 \leqq \nu \leqq \tfrac{1}{2}).$$

We use the following lemma, which is related to the well-posedness of the Stokes problem. For the proof see [12].

LEMMA A.1. *There exists $\underline{v}_0 \in U$ and a constant C_1 depending on ω and $\bar{h}$ such that the following inequalities hold*:

$$\|\underline{v}_0\|_{1,\Omega} \leqq C_1 \|\operatorname{tr}(\underline{\underline{\sigma}})\|_{0,\Omega},$$

$$(\operatorname{div} \underline{v}_0, \operatorname{tr}(\underline{\underline{\sigma}})) \geqq \|\operatorname{tr}(\underline{\underline{\sigma}})\|_{0,\Omega}^2.$$

With $\underline{v}_0$ as in Lemma A.1 we now set $(\underline{v}, \underline{\underline{\tau}}) = (-\underline{u} - \delta \underline{v}_0, \underline{\underline{\sigma}} - \delta^2 \underline{\underline{\varepsilon}}(\underline{u}))$, where δ is a constant to be specified shortly. Then by (A4), the inequality $(\tau_1, \tau_2)_{\mathcal{H}} \leqq (s/2)\|\tau_1\|_{\mathcal{H}}^2 + (1/2s)\|\tau_2\|_{\mathcal{H}}^2$ $(s > 0)$, and Lemma 3.1, we have that

$$\begin{aligned}
\mathcal{B}(\underline{u}, \underline{\underline{\sigma}}; \underline{v}, \underline{\underline{\tau}}) &= (\underline{\underline{\sigma}}, S^{-1}\underline{\underline{\sigma}})_{\mathcal{H}} + \tfrac{1}{3}\delta(\operatorname{tr}(\underline{\underline{\sigma}}), \operatorname{div} \underline{v}_0) \\
&\quad + \delta(\underline{\underline{\sigma}} - \underline{\underline{\sigma}}_0, \underline{\underline{\varepsilon}}(\underline{v}_0)) + \delta^2\|\underline{\underline{\varepsilon}}(\underline{u})\|_{\mathcal{H}}^2 - \delta^2(\underline{\underline{\sigma}}, S^{-1}\underline{\underline{\varepsilon}}(\underline{u}))_{\mathcal{H}} \\
&\geqq (\tfrac{1}{12}\bar{h}^3 - C_2\delta - C_3\delta^2)\|\underline{\underline{\sigma}} - \underline{\underline{\sigma}}_0\|_{\mathcal{H}}^2 + (\tfrac{1}{6}\delta - C_4\delta^2)\|\operatorname{tr}(\underline{\underline{\sigma}})\|_{0,\Omega}^2 + c_1\delta^2\|\underline{u}\|_{1,\Omega}^2 \\
&\geqq \min\{\tfrac{1}{12}\bar{h}^3 - C_2\delta - C_3\delta^2, \tfrac{1}{2}\delta - 3C_4\delta^2, c_1\delta^2\}(\|\underline{u}\|_{1,\Omega}^2 + \|\underline{\underline{\sigma}}\|_{\mathcal{H}}^2).
\end{aligned}$$

Thus, choosing δ to be a sufficiently small positive number, we have found $(\underline{v}, \underline{\tau}) \in U \times \mathcal{H}$ such that $\|\underline{v}, \underline{\tau}\|_H \leqq C\|\underline{u}, \underline{\varrho}\|_H$ and $\mathcal{B}(\underline{u}, \underline{\varrho}; \underline{v}, \underline{\tau}) \geqq c\|\underline{u}, \underline{\varrho}\|_H^2$, where C and c depend only on ω and $\bar{h}$. Hence (A2) is true in case (a) with c depending on ω and $\bar{h}$.

(b) The Reissner-Mindlin model. Given $(w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}) \in H_0^1(\omega) \times V \times \mathcal{H} \times [L_2(\omega)]^2$, let $(z, \underline{\varphi}, \underline{\underline{k}}, \underline{\zeta}) = (-w, -\underline{\theta}, \underline{\underline{m}} - \delta \underline{\underline{\varepsilon}}(\underline{\theta}), \underline{\gamma} - \delta(\underline{\theta} - \underline{\nabla} w))$, where δ is a constant to be specified. Then noting that by (3.6), $(\underline{\underline{m}}, T^{-1}\underline{\underline{m}}) \geqq \|\underline{\underline{m}}\|_{\mathcal{H}}^2/(1+\nu)$, and recalling Lemma 3.3, we have

$$\begin{aligned}
\mathcal{B}(w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}; z, \underline{\varphi}, \underline{\underline{k}}, \underline{\zeta}) &= (\underline{\underline{m}}, T^{-1}\underline{\underline{m}})_{\mathcal{H}} + \left(\frac{h^2}{\kappa}\right)\|\underline{\gamma}\|_{0,\omega}^2 + \delta\|\underline{\underline{\varepsilon}}(\underline{\theta})\|_{\mathcal{H}}^2 - \delta(\underline{\underline{m}}, T^{-1}\underline{\underline{\varepsilon}}(\underline{\theta}))_{\mathcal{H}} \\
&\quad + \delta\|\underline{\theta} - \underline{\nabla} w\|_{0,\omega}^2 - \delta\left(\frac{h^2}{\kappa}\right)(\underline{\gamma}, \underline{\theta} - \underline{\nabla} w) \\
&\geqq \left(\frac{1}{1+\nu} - C_1\delta\right)\|\underline{\underline{m}}\|_{\mathcal{H}}^2 + \frac{1}{2}\delta\|\underline{\underline{\varepsilon}}(\underline{\theta})\|_{\mathcal{H}}^2 + \frac{1}{2}\delta\|\underline{\theta} - \underline{\nabla} w\|_{0,\omega}^2 \\
&\quad + \left(\frac{h^2}{\kappa}\right)\left(1 - C_2\delta\frac{h^2}{\kappa}\right)\|\underline{\gamma}\|_{0,\omega}^2 \\
&\geqq \min\left\{\frac{1}{1+\nu} - C_1\delta, c_1\delta, \frac{h^2}{\kappa}\left(1 - C_2\delta\frac{h^2}{\kappa}\right)\right\}\|w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}\|_H^2.
\end{aligned}$$

Thus if δ is small enough we have found $(z, \underline{\varphi}, \underline{\underline{k}}, \underline{\zeta}) \in H$ such that $\|z, \underline{\varphi}, \underline{\underline{k}}, \underline{\zeta}\|_H \leqq C\|w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}\|_H$ and $\mathcal{B}(w, \underline{\theta}, \underline{\underline{m}}; z, \underline{\varphi}, \underline{\underline{k}}, \underline{\zeta}) \geqq c\|w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}\|_H^2$, where the constants depend only on ω, $\bar{h}$, and $\bar{\kappa}$. These prove the assertion in case (b).

(c) The Kirchhoff model. Given $(w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}) \in H$, let $(z, \underline{\varphi}, \underline{\underline{k}}, \underline{\zeta}) = (-w, -\underline{\theta} - \delta\underline{\varphi}_0, \underline{\underline{m}} - \delta\underline{\underline{\varepsilon}}(\underline{\theta}), \underline{\gamma} - \delta\underline{\zeta}_0)$, where $\underline{\varphi}_0 \in V$ and $\underline{\zeta}_0 \in V'$ are defined so as to satisfy

$$\begin{aligned}
&\|\underline{\varphi}_0\|_{1,\omega} = \|\underline{\gamma}\|_{V'}, \qquad \langle \underline{\gamma}, \underline{\varphi}_0\rangle = \|\underline{\gamma}\|_{V'}^2, \\
&\|\underline{\zeta}_0\|_{V'} = \|\underline{\theta} - \underline{\nabla} w\|_{1,\omega}, \qquad \langle \underline{\theta} - \underline{\nabla} w, \underline{\zeta}_0\rangle = \|\underline{\theta} - \underline{\nabla} w\|_{1,\omega}^2,
\end{aligned}$$

which obviously is possible. As in case (b), we then find that for a sufficiently small δ, $\|z, \underline{\varphi}, \underline{\underline{k}}, \underline{\zeta}\|_H \leqq C\|w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}\|_H$ and $\mathcal{B}(w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}; z, \underline{\varphi}, \underline{\underline{k}}, \underline{\zeta}) \geqq c\|w, \underline{\theta}, \underline{\underline{m}}, \underline{\gamma}\|_H^2$, where C and c depend only on ω, and so the assertion follows in case (c).

Appendix B. The Korn inequality. Let ω be a bounded Lipschitz domain and define the seminorm

$$|\underline{\theta}|_{E(\omega)} = \left\{\int_\omega \sum_{i,j=1}^{2} |\varepsilon_{ij}(\underline{\theta})|^2 \, dx_1 \, dx_2\right\}^{1/2}, \qquad \underline{\theta} \in (H^1(\omega))^2,$$

where $\varepsilon_{ij}(\underline{\theta}) = \frac{1}{2}(\partial\theta_i/\partial x_j + \partial\theta_j/\partial x_i)$, and let

$$|\underline{u}|_{R,\omega}^2 = |\underline{\theta}|_{E(\omega)}^2 + \|\underline{\theta} - \underline{\nabla} w\|_{0,\omega}^2, \quad \underline{u} = (w, \underline{\theta}), \quad w \in H^1(\omega), \quad \underline{\theta} \in (H^1(\omega))^2.$$

THEOREM B.1. *There is a constant C depending only on ω such that for any $\underline{\theta} \in [H^1(\omega)]^2$*

$$\text{(B1)} \qquad \inf_{abc}\{\|\theta_1 - a - bx_2\|_{1,\omega}^2 + \|\theta_2 - c + bx_1\|_{1,\omega}^2\} \leqq C|\underline{\theta}|_{E(\omega)}^2,$$

$$\text{(B2)} \qquad \inf_{abcd}\|w - (a + bx_1 + cx_2 + dx_1x_2)\|_{H^1(\omega)} \leqq C|\underline{u}|_{R,\omega}.$$

Proof. Inequality (B1) follows immediately from the Korn inequality for plane elasticity (see [19]). Inequality (B2) follows from (B1).

LEMMA B.1. *There exists a constant C depending on ω such that for any $(w, \underline{\theta}) \in [H^1(\omega)]^3$*

$$\|w\|^2_{1,\omega} + \|\underline{\theta}\|^2_{1,\omega} \leqq C\left\{|\underline{u}|^2_{R,\omega} + \int_{\partial\omega} w^2\, ds\right\}.$$

Proof. We apply the standard contradiction argument. If the assertion is not true, there is a sequence $\{w_n, \underline{\theta}_n\}$ such that

$$\|w_n, \underline{\theta}_n\|_{1,\omega} = 1,$$

$$\|\underline{\theta}_n\|_{E(\omega)} \to 0,$$

$$\|\underline{\theta}_n - \underline{\underline{\nabla}} w_n\|_{0,\omega} \to 0, \qquad \int_{\partial\omega} w_n^2\, ds \to 0$$

as $n \to \infty$. Then by Theorem B.1, $\{\underline{\theta}_n\}$ contains a subsequence (which we denote once more by $\{\underline{\theta}_n\}$) such that $\underline{\theta}_n \to (a - bx_2, c + bx_1)$ in $[H^1(\omega)]^2$. Furthermore, since $\|\theta_n - \nabla w_n\|_{0,\omega} \to 0$, there is another subsequence (once more denoted by $\{\underline{\theta}_n, w_n\}$) so that $w_n \to w$ in $H^1(\omega)$. Hence $b = 0$ and $w = ax_1 + cx_2 + d$. Because $\int w_n^2\, ds \to 0$ we get $a = c = d = 0$, contradicting the assumption $\|w_n, \underline{\theta}_n\|_{1,\omega} = 1$. □

We immediately get Theorem B.2.

THEOREM B.2. *There exists a constant C depending only on ω such that for any $\underline{u} = (w, \underline{\theta}) \in H^1_0(\omega) \times [H^1(\omega)]^2$*

$$\|w\|^2_{1,\omega} + \|\underline{\theta}\|_{1,\omega} \leqq C|\underline{u}|^2_{R,\omega}. \tag{B3}$$

Let us now consider a family $\mathcal{F} = \{\omega\}$ of Lipschitz bounded domains. The family will be called *regular* if there is a (uniform) constant C so that (B3) holds for all $\omega \in \mathcal{F}$.

Let us now consider a special family of domains. Let $\omega^{[0]}$ be a unit circle and $\omega^{[n]}$ be a sequence of regular $n+3$-polygons such that

$$\bar{\omega}^{[n]} \subset \bar{\omega}^{[n+1]} \subset \bar{\omega}^{[n+1]} \subset \omega^{[0]},$$

$$\omega^{[n]} \to \omega^{[0]} \quad \text{as } n \to \infty$$

in the sense that for any $x \in \omega^{[0]}$ there is $n(x) > 0$ such that $x \in \omega^{[n]}$ for all $n > n(x)$. We let $\mathcal{F}_0 = \{\omega^{[0]}, \omega^{[1]}, \omega^{[2]}, \cdots\}$.

THEOREM B.3. *The family $\mathcal{F}_0$ is a regular family of domains and hence there exists $C > 0$ such that*

$$\|w\|^2_{1,\omega^{[n]}} + \|\underline{\theta}\|^2_{1,\omega^{[n]}} \leqq C|\underline{u}|^2_{R,\omega^{[n]}}$$

for any $\underline{u} = (w, \underline{\theta}) \in H^1_0(\omega^{[n]}) \times [H^1(\omega^{[n]})]^2$, $n = 0, 1, 2, \cdots$.

Proof. For $n > n_0$ the $\omega^{[n]}$ are star-shaped domains and

$$\partial\omega^{[n]} = \{(x_1, x_2): x_1 = \rho_n(\theta)\cos\theta, x_2 = \rho_n(\theta)\sin\theta, 0 \leqq \theta \leqq 2\pi\},$$

where $\rho_n(\theta) \to 1$ and $\rho_n'(\theta) \to 0$ uniformly. Let Q_n be the one-to-one map of $\omega^{[n]}$ onto $\omega^{[0]}$ defined by

$$Q_n(\rho(\theta)\cos\theta, \rho(\theta)\sin\theta) = (\rho(\theta)\cos\theta, \rho(\theta)\sin\theta) \quad \text{for } \rho(\theta) \leqq \frac{1}{2}$$

$$= \left(\frac{1}{2}\frac{\rho(\theta) - (\frac{1}{2})}{\rho_n(\theta) - (\frac{1}{2})} + \frac{1}{2}\right)\cos\theta$$

$$= \left(\frac{1}{2}\frac{\rho(\theta) - (\frac{1}{2})}{\rho_n(\theta) - (\frac{1}{2})} + \frac{1}{2}\right)\sin\theta \quad \text{for } \rho(\theta) > \frac{1}{2}.$$

If $Q_n(x_1, x_2) = (\xi_1, \xi_2)$ then we have $\xi_1 = \xi_1^{[n]}(x_1, x_2)$, $\xi_2^{[n]} = \xi_2(x_1, x_2)$, $x_1 = x_1^{[n]}(\xi_1, \xi_2)$, $x_2 = x_2^{[n]}(\xi_1, \xi_2)$, and $\xi_i^{[n]} \to x_i$, $(\partial\xi_i/\partial x_j) \to \delta_{ij}$, $x_i^{[n]} \to \xi_i$, $(\partial x_i/\partial\xi_j) \to \delta_{ij}$, $i, j = 1, 2$ as $n \to \infty$, uniformly with respect to $(x_1, x_2) \in \omega^{[n]}$ and $(\xi_1, \xi_2) \in \omega^{[0]}$. Let $\underline{u} = (w, \underline{\theta}) \in H_0^1(\omega^{[n]}) \times (H^1(\omega^{[n]}))^2$ and let

$$\bar{u} = (\bar{w}, \bar{\underline{\theta}}), \qquad \bar{\underline{u}}(\xi_1, \xi_2) = \underline{u}(x_1(\xi_1, \xi_2), x_2(\xi_1, \xi_2)).$$

Then $\bar{\underline{u}} \in H_0^1(\omega^{[0]}) \times (H^1(\omega^{[0]})^2)$ and by Theorem B.2 we have

$$\|\bar{w}\|_{1,\omega^{[0]}}^2 + \|\underline{\theta}\|_{1,\omega^{[0]}}^2 \leqq C|\bar{\underline{u}}|_{R,\omega^{[0]}}^2$$

and also

$$\|\bar{w}\|_{1,\omega^{[0]}} = \|w\|_{1,\omega^{[n]}}(1 + o(1)),$$

$$\|\bar{\underline{\theta}}\|_{1,\omega^{[0]}} = \|\underline{\theta}\|_{1,\omega^{[n]}}(1 + o(1)),$$

$$|\bar{\underline{u}}|_{R,\omega^{[0]}} = |\underline{u}|_{R,\omega^{[n]}} + o(1)(\|w\|_{1,\omega^{[n]}} + \|\underline{\theta}\|_{1,\omega^{[n]}})$$

as $n \to \infty$. Hence

$$\|w\|_{1,\omega^{[n]}}^2(1 + o(1)) + \|\underline{\theta}\|_{1,\omega^{[n]}}^2(1 + o(1)) \leqq C[|\underline{u}|_{R,\omega^{[n]}}^2 + o(1)(\|w\|_{1,\omega^{[n]}}^2 + \|\underline{u}\|_{1,\omega^{[n]}}^2)].$$

From this we see that for $n > n_0$ the family is a regular one. Using Theorem B.2, we then see that the whole family $\mathscr{F}_0$ is regular. □

Appendix C. A projection theorem.

THEOREM C.1. *Let H be a Hilbert space, let $\{H_n\}$ and $\{K_n\}$ be sequences of closed subspaces of H such that $H_n \subset H_{n+1}$ and $K_n \supseteq K_{n+1}$, $n = 1, 2, \cdots$, and let*

$$H_0 = \overline{\bigcup_n H_n} \quad \text{and} \quad K_0 = \bigcap_n K_n.$$

Furthermore, let P_n and Q_n, respectively, P_0, Q_0, be orthogonal projections onto H_n and K_n, respectively, H_0, K_0. Then for any $u \in H$

$$\|P_n u - P_0 u\| \to 0, \qquad \|Q_n u - Q_0 u\| \to 0$$

as $n \to \infty$.

Proof. First observe that $\|Q_{n+1}u\| = \|Q_{n+1}Q_n u\| \leqq \|Q_n u\|$, so $\|Q_n u\| \to q \geqq 0$ monotonically. Furthermore,

$$\|Q_n u - Q_{n+j}u\| = \|Q_n u\|^2 - 2(Q_n u, Q_{n+j}u) + \|Q_{n+j}\|^2 = \|Q_n u\|^2 - \|Q_{n+j}u\|^2,$$

so $\{Q_n u\}$ is a Cauchy sequence. So $Q_n u \to v$ and $v \in K_n$ for all n. Hence $v \in K_0$ and since $(v, w) = \lim_{n\to\infty}(Q_n u, w) = \lim_{n\to\infty}(u, Q_n w) = (u, w)$ for all $w \in K_0$, it follows that $v = Q_0 u$.

Let us now consider the projection operator $I - P_n = \tilde{Q}_n$. Then $\tilde{Q}_n$ projects H onto H_n^1 and $H_w^1 \supset H_{n+1}^1$. Hence $\tilde{Q}_n u = u - P_n u \to u - v \in \bigcap H_n^1$. So $P_n u \to v \in H_0$ and by the same argument as before, $v = P_0 u$. □

REFERENCES

[1] S. A. AMBARCUMJAN, *Theory of Anisotropic Plates*, Second edition, Nauka, Moscow, 1987. (In Russian.)

[2] I. BABUŠKA, *Stability of the domain under perturbation of the boundary in fundamental problems in the theory of partial differential equations principally in connection with the theory of elasticity* 1, 2, Czechoslovak Math. J., 11 (1961), pp. 75-105, 165-203. (In Russian.)

[3] ———, *The stability of domains and the question of the formulation of the plate problem*, Apl. Mat., (1962), pp. 463-467.

[4] I. BABUŠKA, *The theory of small changes in the domain of existence in the theory of partial differential equations and its applications*, in Differential Equations and Their Applications, Proc. Conference, Prague, 1962, Academic Press, New York, 1963, pp. 13-26.

[5] I. BABUŠKA AND A. K. AZIZ, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Application to Partial Differential Equations, A. K. Aziz, ed., Academic Press, New York, 1973, pp. 5-363.

[6] I. BABUŠKA, *Uncertainties in engineering design: mathematical theory and numerical experience*, in Optimal Shape—Automated Structural Design, J. Bennett and M. Botkin, eds., Plenum Press, New York, 1986.

[7] I. BABUŠKA AND T. SCAPOLLA, *Benchmark computation and performance evaluation for a rhombic plate bending problem*, Internat. J. Numer. Methods Engrg., 28 (1989), pp. 155-181.

[8] G. BIRKHOFF, *The Numerical Solution of Elliptic Equations*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1971.

[9] F. BREZZI, *On the existence, uniqueness and approximation of saddle point problems arising from Lagrange multipliers*, RAIRO, Ser R., 8 (1974), pp. 129-151.

[10] P. G. CIARLET AND P. DESTUYNDER, *A justification of the two-dimensional linear plate model*, J. Méc. Théor. Appl., 18 (1979), pp. 315-344.

[11] P. DESTUYNDER, *Une théorie asymptotique des plaques minces en élasticité linéaire*, Masson, Paris, New York, 1986.

[12] V. GIRAULT AND P. A. RAVIART, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer-Verlag, Berlin, New York, 1986.

[13] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[14] V. G. MAZ'YA AND S. A. NAZAROV, *About the Sapondzhyn-Babuška paradox in the plate theory*, Dokl. Akad. Nauk Armenian Rep., 78 (1984), pp. 127-130. (In Russian.)

[15] ———, *Paradoxes of limit passage in solutions of boundary value problems involving the approximation of smooth domains by polygonal domains*, Izv. Akad. Nauk SSSR Ser. Mat., 50 (1986), pp. 1156-1177. (In Russian.) Math. USSR-Izv., 29 (1987), pp. 511-533. (In English.)

[16] D. MORGENSTERN, *Herleitung der Plattentheorie aus der dreidimensionalen Elastizitätstheorie*, Arch. Rational Mech. Anal., 4 (1959), pp. 145-152.

[17] D. MORGENSTERN AND I. SZABÓ, *Vorlesungen über theoretische Mechanik*, Springer-Verlag, Berlin, New York, 1961.

[18] N. W. MURRAY, *The polygon-circle paradox in thin plate theory*, Proc. Cambridge Philos. Soc., 73 (1973), pp. 279-282.

[19] J. NEČAS AND I. HLAVACEK, *Mathematical Theory of Elastic and Elasto-Plastic Bodies: An Introduction*, Elsevier, Amsterdam, New York, 1981.

[20] J. PITKÄRANTA, *Analysis of some low-order finite element schemes for Mindlin-Reissner and Kirchhoff plates*, Numer. Math., 53 (1988), pp. 237-254.

[21] K. RAJAIAH AND A. K. RAO, *On the polygon-circle paradox*, Trans. ASME Ser. E., J. Appl. Mech., 48 (1981), pp. 195-196.

[22] E. REISSNER, *Reflections on the theory of elastic plates, developments in mechanics*, Appl. Mech. Rev., 38 (1985), pp. 1453-1464.

[23] G. RIEDER, *On the plate paradox of Sapondzhyan and Babuška*, Mech. Res. Comm., 1 (1974), pp. 51-53.

[24] A. S. SAADA, *Elasticity Theory and Applications*, Pergamon Press, New York, 1974.

[25] O. M. SAPONDZHYAN, *Bending of a freely supported polygonal plate*, Izv. Akad. Nauk Armyan SSR Ser. Fiz-Mat. Estestv. Tekhn. Nauk, 5 (1952), pp. 29-46.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 577-592, May 1990
© 1990 Society for Industrial and Applied Mathematics
002

A BIDIMENSIONAL ELECTROMAGNETIC PROBLEM*

MICHEL CROUZEIX† AND JEAN DESCLOUX‡

Abstract. Electric alternating currents running in a system of infinite cylindrical parallel conductors generate an electromagnetic field, which is defined by a potential $\varphi : \mathbb{R}^2 \to \mathbb{C}$. φ is harmonic in the exterior domain and satisfies a Helmholtz equation in the cross sections of the conductors. The asymptotic properties of φ as the frequency tends to infinity, in particular the boundary layers due to the skin effect, are studied.

Key words. electromagnetic fields, asymptotic behaviour, boundary layer, skin effect, Helmholtz equation

AMS(MOS) subject classifications. 35J05, 35B25, 35B40

1. Introduction. This paper has been motivated by the numerical simulation of the electromagnetic casting process (EMC). When the ingot is sufficiently long, the electromagnetic part of the problem reduces to the search of a complex potential φ in $\mathbb{R}^2$, of class C^1, with logarithmic behaviour at infinity, satisfying the conditions

$$\Delta\varphi + 2i\alpha^2(\varphi + C_k) = 0 \quad \text{in } \Omega_k, \quad 1 \leqq k \leqq N, \qquad \Delta\varphi = 0 \quad \text{in } \bar{\Omega}^c; \tag{1.1}$$

here $\Omega_k \subset \mathbb{R}^2$, $1 \leqq k \leqq N$, are the cross sections in the x_1, x_2 plane of cylindrical electric conductors in which runs a current with angular frequency ω; $2\alpha^2 = \mu_0\sigma\omega$ is a real constant where μ_0 is the magnetic permeability of the air and σ the conductivity; i is the imaginary unit and the C_k's are given complex constants.

Our initial purpose was to find a fast, accurate, and reliable numerical algorithm for solving problem (1.1). We have been led to a theoretical study of problem (1.1), partly independent of our original aim, which is the subject of this paper. We shall report on the numerical aspects in a forthcoming paper.

Maybe because of the problem's particular character, we have found few references in the literature relative to problem (1.1). All of them [2], [4] are essentially concerned with numerical, practical, or theoretical questions.

In § 2, we introduce some notation and recall a few basic mathematical tools and results. We feel it is necessary to construct with some detail our mathematical model; this is the object of § 3. In § 4, because of their physical and mathematical relevance, we define two specific problems, closely related to (1.1), which are studied in the remaining parts of the paper.

For many applications, in particular for the EMC process, the coefficient α is large and we can observe a very pronounced skin effect in the conductors, i.e., the electric current $j = 2i\alpha^2(\varphi + C_k)$ almost vanishes in Ω_k except in a thin layer in the neighborhood of $\partial\Omega_k$. This is a serious difficulty for numerical computations, in particular, since we are specially interested in the fields (current, induction, Laplace forces) in the conductors. For this reason in §§ 5-7 we analyse the behaviour of φ for large α. Section 5 gives some global estimates for the rate of convergence of φ to a limit function φ_∞, as α tends to infinity. Section 6 contains more refined local estimates. In § 7 we construct a simple and accurate boundary layer approximation of φ in the neighborhood of smooth parts of $\partial\Omega_k$. All these results are rather satisfactory and complete except in the neighborhoods of corners of $\partial\Omega_k$ for which many improvements

* Received by the editors February 29, 1988; accepted for publication (in revised form) May 26, 1989.

† Mathématiques, Institut de Recherche en Informatique et Systèmes Aléatoires, Université de Rennes, Campus de Beaulieu, 35042 Rennes Cedex, France.

‡ Département de Mathématiques, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland.

remain to be done; in particular, we have not succeeded in getting a simple picture of the singularity of φ at a corner for large α.

As a consequence of Proposition 7.2, it follows that, along regular parts of $\partial\Omega_k$, φ satisfies approximatively a Robin-type boundary condition. In § 8 we show how this property can be used, in the case where the $\partial\Omega_k$'s have no corner, for defining an approximation of φ, which can be computed easily.

2. Notation and some basic tools. An element of $\mathbb{R}^2$ is denoted by $x=(x_1, x_2)$; $|x|^2=x_1^2+x_2^2$. $B(x,\delta)$ is the open ball with center x and radius δ. ∂_1 and ∂_2 are the partial derivatives with respect to x_1 and x_2 and $\vec{\nabla}=(\partial_1,\partial_2)$. For $m\geqq 0$, ∂_1^m and ∂_2^m are the mth derivatives with respect to x_1 and x_2. For $\Lambda\subset\mathbb{R}^2$, $\bar{\Lambda}$ is the closure of Λ, $\partial\Lambda$ is the boundary of Λ and χ_Λ is the characteristic set of Λ. For a function $v:\Lambda\to\mathbb{C}$, $\bar{v}$ is the complex conjugate of v, and $\partial v/\partial n$ denotes the exterior normal derivative. For $\Lambda\subset\mathbb{R}^2$ open, the symbols $L^p(\Lambda)$, $H^m(\Lambda)$, $W^{m,p}(\Lambda)$ (Sobolev spaces), $C^m(\Lambda)$, $C_0^m(\Lambda)$, $C^{m,\beta}(\Lambda)$ (the mth derivatives are β Hölder continuous, $0<\beta<1$) and the corresponding norms $\|\cdot\|_{L^p(\Lambda)},\cdots$, have their usual meanings for complex-valued functions. We will also use the spaces $L^p_{\mathrm{loc}}(\Lambda),\cdots$ and $L^p(\partial\Lambda)\cdots$.

For an open set $\Lambda\subset\mathbb{R}^2$, $\partial\Lambda$ is piecewise C^∞ if $\partial\Lambda$ is composed of a finite number of closed arcs that are C^∞ and if for each interior angle γ at corners we have $0<\gamma<2\pi$; furthermore, we impose the condition $\partial\Lambda=\partial\bar{\Lambda}$.

Let $\Lambda\subset\mathbb{R}^2$ be an open connected set such that Λ^c is bounded and let $\Phi\neq\varnothing$ be an open bounded subset of Λ. We define

$$W_0^1(\Lambda)=\left\{v:\Lambda\to\mathbb{C}\;\middle|\;\frac{v}{(1+|x|)\ln(2+|x|)}\in L^2(\Lambda):\partial_l v\in L^2(\Lambda),\ l=1,2\right\}, \tag{2.1}$$

$$\|v\|^2_{W_0^1(\Lambda)}=\|v\|^2_{L^2(\Phi)}+\|\partial_1 v\|^2_{L^2(\Lambda)}+\|\partial_2 v\|^2_{L^2(\Lambda)}. \tag{2.2}$$

For a study of the space $W_0^1(\Lambda)$, we can consult, for example, [6]. In particular, we have the following proposition.

PROPOSITION 2.1. (a) *$W_0^1(\Lambda)$ is a complete space for the norm $\|\cdot\|_{W_0^1(\Lambda)}$.*

(b) *$C_0^\infty(\mathbb{R}^2)$ is dense in $W_0^1(\mathbb{R}^2)$.*

(c) *If $\partial\Lambda$ is piecewise C^∞, then $W_0^1(\Lambda)=\{v_{|\Lambda}\mid v\in W_0^1(\mathbb{R}^2)\}$.*

(d) $1\in W_0^1(\Lambda)$, $\ln(1+|x|)\notin W_0^1(\mathbb{R}^2)$.

The following result of potential theory is classical.

PROPOSITION 2.2. *Let v be an harmonic complex function on the open domain $\Lambda\subset\mathbb{R}^2$ with bounded complement. We suppose that $v(x)=O(\ln|x|)$ as $|x|\to\infty$. Then there exist complex constants d and e such that*

$$v(x)=d\ln|x|+e+O(|x|^{-1}),$$
$$\partial_1^k\partial_2^l v(x)=d\,\partial_1^k\partial_2^l\ln|x|+O(|x|^{-(k+l+1)}),\qquad k+l\geqq 1.$$

The next trace estimate will play an important role in § 5.

PROPOSITION 2.3. *Let $\Lambda\subset\mathbb{R}^2$ be an open domain such that $\partial\Lambda$ is piecewise C^∞. Then there exists a constant c independent of $v\in H^1(\Lambda)$ such that*

$$\|v\|^2_{L^2(\partial\Lambda)}\leqq c\|v\|_{H^1(\Lambda)}\|v\|_{L^2(\Lambda)}. \tag{2.3}$$

Proof. By a classical imbedding theorem (see, for example, [1]), for any $w\in L^1(\Lambda)$ such that $\partial_l w\in L^1(\Lambda)$, $l=1,2$, we have $\|w\|_{L^1(\partial\Lambda)}\leqq c\{\|w\|_{L^1(\Lambda)}+\|\partial_1 w\|_{L^2(\Lambda)}+\|\partial_2 w\|_{L^1(\Lambda)}\}$, where c is some generic constant. We apply this result to $w=v^2$ and by Schwarz's inequality we obtain

$$\begin{aligned}(\|v\|_{L^2(\partial\Lambda)})^2&\leqq c\{\|v\|^2_{L^2(\Lambda)}+\|v\|_{L^2(\Lambda)}(\|\partial_1 v\|_{L^2(\Lambda)}+\|\partial_2 v\|_{L^2(\omega)})\}\\&\leqq c\|v\|_{L^2(\Lambda)}\|v\|_{H^1(\Lambda)}.\end{aligned}$$

□

3. Mathematical model. In $\mathbb{R}^3$ we consider a system of N cylindrical conductors parallel to the x_3 axis in an alternating electromagnetic field. We denote by $\Omega_1, \Omega_2, \cdots, \Omega_N \subset \mathbb{R}^2$ their cross sections, which we suppose are bounded in the x_1, x_2 plane, and set $\Omega = \cup\, \Omega_k$. Let $\vec{j}$, E, and H be, respectively, the tridimensional complex current, the electric field, and the magnetic field. We introduce the following assumptions:

(3.1) (a) $\vec{E}(x_1, x_2, x_3, t) = E(x_1, x_2)\, e^{-i\omega t}\, \vec{e}_3$ in the conductors,

(3.2) (b) $\vec{H}(x_1, x_2, x_3, t) = H_1(x_1, x_2)\, e^{-i\omega t}\, \vec{e}_1 + H_2(x_1, x_2)\, e^{-i\omega t}\, \vec{e}_2$ in $\mathbb{R}^3$,

(3.3) (c) $\vec{j} = \sigma \vec{E}$ in the conductors, $\vec{j} = 0$ outside the conductors,

(3.4) (d) $\partial_t \vec{H} + 1/\mu_0 \overrightarrow{\text{rot}}\, E = 0$ in the conductors,

(3.5) (e) $\overrightarrow{\text{rot}}\, \vec{H} = \vec{j}$ in $\mathbb{R}^3$,

(3.6) (f) $\text{div}\, \vec{H} = 0$ in $\mathbb{R}^3$.

Let us be precise and comment on these hypotheses. i is the complex unit; ω is the angular frequency; t is the time; μ_0 is the magnetic permeability in the vacuum; $\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$ are the unit vectors along the coordinate axis. Assumption (3.1) is assumed to be valid only in the conductors; supposing (3.1) is valid in $\mathbb{R}^3$ is physically not realistic (see Sommerfeld [7]) and would lead to mathematical contradictions. In (3.3), we suppose that the electric conductivity σ is the same positive constant for all conductors; in fact, all results contained in this paper can be extended without difficulty to the case where the conductivity is constant in each conductor but different from one to another. Assumptions (3.4)-(3.6) are standard Maxwell's equations where in (3.5) we have neglected the displacement currents, which is legitimate for moderate frequencies. Finally, we observe that all fields are x_3 independent and can be considered as mapping from $\mathbb{R}^2$ into $\mathbb{C}$ or $\mathbb{C}^2$.

From (3.1), (3.3), there exists $j : \mathbb{R}^2 \to \mathbb{C}$ such that

$$\begin{aligned} &\vec{j}(x_1, x_2, x_3, t) = j(x_1, x_2)\, e^{-i\omega t}\, \vec{e}_3, \quad \text{and} \\ &j(x) = \sigma E(x) \quad \text{for } x \in \Omega, \qquad j(x) = 0 \quad \text{for } x \in \Omega^c, \end{aligned} \tag{3.7}$$

where $x = (x_1, x_2) \in \mathbb{R}^2$. From (3.6) and (3.2) follows the existence of a potential φ in $\mathbb{R}^2$ such that

$$(H_1(x), H_2(x)) = \overrightarrow{\text{rot}}\, \varphi(x) = (\partial_2 \varphi(x), -\partial_1 \varphi(x)), \qquad x \in \mathbb{R}^2 \tag{3.8}$$

where $\overrightarrow{\text{rot}}$ denotes here the bidimensional vector curl. By using in particular (3.1), (3.4), (3.7), and (3.8), we obtain

$$\overrightarrow{\text{rot}}\, (i\omega\mu_0\sigma\varphi - j) = 0 \quad \text{in } \Omega_k, \quad k = 1, 2, \cdots, N;$$

this implies the existence of constants $C_k \in \mathbb{C}$ such that

$$j = i\omega\mu_0\sigma(\varphi + C_k) \quad \text{in } \Omega_k, \quad k = 1, 2, \cdots, N. \tag{3.9}$$

By (3.2), (3.5), (3.7), and (3.8), we have

$$-\Delta\varphi = j \quad \text{in } \mathbb{R}^2; \tag{3.10}$$

finally, with (3.7), (3.9), we obtain

$$-\Delta\varphi = i\omega\mu_0\sigma\chi\varphi + f \quad \text{in } \mathbb{R}^2 \tag{3.11}$$

where χ is the characteristic function of Ω and $f = i\omega\mu_0\sigma C_k$ in Ω_k, $f = 0$ in Ω^c.

For physical reasons, it is natural to impose on φ a behaviour at most logarithmic at infinity. By Proposition 2.2 there exist constants d and $e \in \mathbb{C}$ such that

$$\varphi(x) = d \ln |x| + e + O(|x|^{-1}) \quad \text{as } |x| \to \infty. \tag{3.12}$$

Without restriction of generality, we can further impose the constraint $e = 0$ since the physical fields are independent of e.

On the basis of equations (3.11), (3.12) with the condition $e = 0$, we are now in the position of formulating the mathematical problem we will study.

Let $\Omega = \bigcup_k \Omega_k \subset \mathbb{R}^2$; we suppose

(3.13) (a) Ω_k is a bounded connected open set, $1 \leqq k \leqq N$;

(3.14) (b) $\bar{\Omega}_k \cap \bar{\Omega}_l = \varnothing$ for $k \neq l$; $\partial\Omega$ is piecewise C^∞.

For given constants $C_k \in \mathbb{C}$, $k = 1, 2, \cdots, N$, our basic mathematical problem will be to find $\varphi \in L^2_{\text{loc}}(\mathbb{R}^2)$ such that

$$-\Delta\varphi = 2i\alpha^2 \chi_\Omega \varphi + f \quad \text{in distribution in } \mathbb{R}^2, \tag{3.15}$$

$$\varphi(x) = d \ln |x| + O(|x|^{-1}) \quad \text{as } |x| \to \infty, \tag{3.16}$$

where χ_Ω is the characteristic function of Ω, $f = 2i\alpha^2 C_k$ in Ω_k, $f = 0$ in Ω^c and $\alpha^2 = \omega\mu_0\sigma/2 > 0$; in (3.16) $d \in \mathbb{C}$ is an unknown constant.

From potential theory (see, for example, Nedelec [6]), it follows that problem (3.15), (3.16) is equivalent to the following integral problem. For given constants $C_k \in \mathbb{C}$, find $\varphi \in L^2(\Omega)$ such that

$$\varphi(x) = \frac{-i\alpha^2}{\pi} \int_\Omega \ln |\xi - x| \varphi(\xi)\, d\xi - \frac{1}{2\pi} \int_\Omega \ln |\xi - x| f(\xi)\, d\xi. \tag{3.17}$$

Our first basic result is contained in the following.

PROPOSITION 3.1. *Problem* (3.15), (3.16) *has one and only one solution. Furthermore,* $\varphi \in W^{2,p}_{\text{loc}}(\mathbb{R}^2)$, *for any* $1 \leqq p < \infty$, $\varphi \in C^{1,\beta}(\mathbb{R}^2)$ *for any* $0 < \beta < 1$, *and* $\varphi \in C^\infty(\Omega \cup \bar{\Omega}^c)$.

Proof. For proving the existence and uniqueness of the solution, it suffices, by the Fredholm alternative, to verify that, for $f = 0$, (3.17) admits only the trivial solution; we multiply (3.17) by $\bar{\varphi}(x)$ and integrate over Ω; we then remark that the left member of the resulting equation is real, whereas the right member is purely imaginary. Standard regularity results for the Laplacian operator show that $\varphi \in C^\infty(\Omega \cup \bar{\Omega}^c)$ and $\varphi \in W^{2,p}_{\text{loc}}(\mathbb{R}^2)$; therefore $\varphi \in C^{1,\beta}_{\text{loc}}(\mathbb{R}^2)$; Proposition 2.2 implies that, in fact, $\varphi \in C^{1,\beta}(\mathbb{R}^2)$, $0 < \beta < 1$. □

For a solution of (3.15), (3.16), we introduce the notion of total current defined by

$$J = \int_\Omega j(x)\, dx. \tag{3.18}$$

Since $2\alpha^2 = \omega\mu_0\sigma$, we obtain by (3.9)

$$J = 2\alpha^2 i \sum_{k=1}^{N} \int_{\Omega_k} (\varphi + C_k)(x)\, dx. \tag{3.19}$$

Furthermore, we have the following relation.

PROPOSITION 3.2. *Let* φ *be the solution of* (3.15), (3.16). *Then*

$$J = -2\pi d.$$

Proof. By Proposition 2.2, we have $d\varphi/dn = d/R + O(R^{-2})$. Integrating relation (3.10) on $B(0, R)$, we obtain the desired result by using Green's formula and (3.18). □

Remark 3.1. In our model, we have neglected the displacement currents, which supposes α is not too large. On another side, most of the results of this paper concern asymptotic relations as $\alpha \to \infty$. Besides their intrinsic mathematical interest, these asymptotic relations are physically relevant in many concrete situations for a large range of values of α.

4. Two particular cases of physical importance. In simulating the industrial process of electromagnetic casting (EMC), we have encountered the following particular case of the situation described in the preceding section. For each $1 \leqq k \leqq N$, there exists $1 \leqq l \leqq N$ such that Ω_k and Ω_l are symmetric with respect to the axis Ox_2; furthermore $C_k = -C_l$; k may be equal to l in which case $C_k = 0$. The constants C_k are obtained from the specifications of the EMC installation. By uniqueness of the solution of problem (3.15), (3.16), it follows that $\varphi(-x_1, x_2) = -\varphi(x_1, x_2)$ for all $(x_1, x_2) \in \mathbb{R}^2$, which implies that $d = 0$ in (3.16). As a consequence of Propositions 2.1, 2.2, and 3.1, $\varphi \in W_0^1(\mathbb{R}^2)$, so that by (3.15) φ satisfies the variational equation

$$\int_{\mathbb{R}^2} \vec{\nabla}\varphi \cdot \vec{\nabla}\bar{v} - 2i\alpha^2 \int_\Omega \varphi\bar{v} = \int_\Omega f\bar{v} \quad \forall v \in W_0^1(\mathbb{R}^2). \tag{4.1}$$

For mathematical convenience, we generalize the above situation somewhat and consider the following problem. For given sets Ω_k satisfying (3.13), (3.14) and for given constants $C_k \in \mathbb{C}$, $1 \leqq k \leqq N$, find $\varphi \in W_0^1(\mathbb{R}^2)$ such that

$$\int_{\mathbb{R}^2} \vec{\nabla}\varphi \cdot \vec{\nabla}\bar{v} - 2i\alpha^2 \int_\Omega \varphi\bar{v} = \int_\Omega f\bar{v} \quad \forall v \in W_0^1(\mathbb{R}^2), \tag{4.2}$$

where $\Omega = \cup_{k=1}^N \Omega_k$, $f = 2i\alpha^2 C_k$ on Ω_k, $f = 0$ in Ω^c.

We will study problem (4.2) in detail in the next sections. However, we immediately remark that, by Propositions 2.1 and 2.2, a solution φ of (4.2) will behave at infinity as $e + O(|x|^{-1})$, and consequently will be a solution of the original problem (3.15), (3.16) if and only if the constant $e = 0$.

Another situation of interest is the case of a single conductor so that $\Omega = \Omega_1$. In this context, however, the total current J defined in (3.18) appears as a data and the constant C_1 as an unknown. By Proposition 3.2, the relevant problem then reads as follows. For given $J \in \mathbb{C}$, find $\varphi \in L^2_{\text{loc}}(\mathbb{R}^2)$ and $\xi \in \mathbb{C}$ such that

$$-\Delta\varphi = 2i\alpha^2 \chi_\Omega(\varphi + \xi) \quad \text{in distribution in } \mathbb{R}^2, \tag{4.3}$$

$$\varphi(x) = \frac{J}{2\pi} \ln|x| + O(|x|^{-1}) \quad \text{as } |x| \to \infty. \tag{4.4}$$

In (4.4), the constant J is immaterial; furthermore, from a mathematical point of view, the number of conductors is of little importance. In the next sections we will consider the following problem. For given Ω_k, $1 \leqq k \leqq N$ satisfying (3.13), (3.14), find $\varphi \in L^2_{\text{loc}}(\mathbb{R}^2)$ and $\xi \in \mathbb{C}$ such that

$$-\Delta\varphi = 2i\alpha^2 \chi_\Omega(\varphi + \xi) \quad \text{in distribution in } \mathbb{R}^2, \tag{4.5}$$

$$\varphi(x) = \ln|x| + O(|x|^{-1}) \quad \text{as } |x| \to \infty. \tag{4.6}$$

We conclude this section by two remarks.

Remark 4.1. The single conductor problem (4.3), (4.4) can be generalized in a natural way for the case of several conductors: we prescribe the current J_k for each of them. By (3.5), this leads to the following problem.

For given $J_k \in \mathbb{C}$, $1 \leqq k \leqq N$, find $\varphi \in L^2_{\text{loc}}(\mathbb{R}^2)$ and the constants $C_k \in \mathbb{C}$, $1 \leqq k \leqq N$, such that (3.15), (3.16) are satisfied under the constraint

$$2i\alpha^2 \int_{\Omega_k} (\varphi + C_k)(x)\, dx = J_k, \qquad 1 \leqq k \leqq N.$$

Clearly, by (3.15), the total current J will be the sum of the J_k. This problem, which is mentioned in the literature [2], possesses one and only one solution. To a large extent its study, in particular the convergence when $\alpha \to \infty$, can be reduced to that of problems (4.2) and (4.5), (4.6). For the sake of simplicity, we will not consider it in the following.

The problems of solving (3.15), (3.16) for given constants C_k on one side and for given J_k on another are strongly related. Suppose we have solved (3.15), (3.16) for $C_j = 1$, $C_l = 0$ if $l \neq j$, $1 \leqq j \leqq N$; then the problem for given J_k reduces to a linear system of N equations that can be proved to possess a unique solution.

Remark 4.2. We can object that the problem introduced in Remark 4.1 is not physical if the total current J does not vanish; in particular, this will always be the case for the situation of the single conductor. The following argument shows that this difficulty can be overcome. To this end, we add to the system an auxiliary conductor Ω_{N+1} for which we prescribe a current equal to $-J$; we then translate it in a given direction. Let b denote the distance between the origin and the center of gravity of Ω_{N+1}, and let φ_b and φ be the solutions of the auxiliary and of the original problems, respectively. It is then possible to show that, for any fixed α and for any fixed bounded domain Λ, we have $\|\varphi_b - \varphi\|_{H^1(\Lambda)} = O(1/b)$.

5. Global asymptotic estimates. As mentioned in the Introduction, we are interested in situations with pronounced skin effect in the conductors. It is then natural to consider what happens if we let the angular frequency ω tend to infinity; since $2\alpha^2 = \mu_0 \sigma \omega$, this amounts to letting the parameter $\alpha \to \infty$. Unfortunately, in problem (3.15), (3.16), for exceptional domains Ω, d may become unbounded as α tends to infinity; see Remark 5.2 below. For this reason, we will treat problems (4.2) and (4.5), (4.6). To insist on the dependence on α, we will sometimes use the notation φ_α, ξ_α, $\cdots$ instead of φ, ξ, $\cdots$. We begin with problem (4.2).

PROPOSITION 5.1. *Problem* (4.2) *has one and only one solution. Furthermore,* $\varphi \in C^{1,\beta}(\mathbb{R}^2) \cap W^{2,p}_{\text{loc}}(\mathbb{R}^2)$, *for all* $0 < \beta < 1$, $1 \leqq p < \infty$, *and there exists a constant* $e \in \mathbb{C}$ *such that*

$$\varphi(x) = e + O(|x|^{-1}) \quad \text{as } |x| \to \infty.$$

Proof. Multiplying (4.2) by $(1+i)$, we see that the left member then defines a continuous and coercive form on $W^1_0(\mathbb{R}^2)$ defined in § 2; existence and uniqueness follows by the Lax-Milgram Lemma. Regularity follows from the fact that φ satisfies (3.15). The asymptotic relation as $|x| \to \infty$ can be deduced from Propositions 2.1 and 2.2. □

We now introduce the function φ_∞, candidate for the limit of φ_α when $\alpha \to \infty$.

LEMMA 5.1. *In connection with problem* (4.2), *there exists a unique function* $\varphi_\infty \in W^1_0(\mathbb{R}^2)$ *such that* $\varphi_\infty = -C_k$ *in* Ω_k *and* $\Delta\varphi_\infty = 0$ *in* $\bar{\Omega}^c$. *Furthermore,* $d\varphi_\infty/dn \in L^2(\partial\Omega)$, *where* $d\varphi_\infty/dn$ *is the normal derivative, exterior to* Ω, *of the restriction of* φ *to* $\bar{\Omega}^c$.

Proof. It suffices to find $\varphi \in W^1_0(\bar{\Omega}^c)$ such that $\Delta\varphi = 0$ and $\varphi = -C_k$ on $\partial\Omega_k$. This problem can be reduced to a problem of the following form. Find $\sigma \in G$ such that $\int_{\bar{\Omega}^c} \nabla\sigma \cdot \nabla\bar{v} = \int_{\bar{\Omega}^c} g\bar{v}$ for all $v \in G$, where g is a bounded function with bounded support and $G = \{v \in W^1_0(\bar{\Omega}^c) \mid v = 0 \text{ on } \partial\Omega\}$; we remark that the Dirichlet form is coercive on

G and apply the Lax-Milgram Lemma. That $d\varphi/dn \in L^2(\partial\Omega)$ is a consequence of the fact that $\partial\Omega$ is piecewise C^∞ (see Grisvard [3]). □

PROPOSITION 5.2. *Let φ_α be the solution of problem* (4.2), *and let φ_∞ be defined by Lemma* 5.1. *Then asymptotically, as $\alpha \to \infty$, we have*

(a) $\|\varphi_\alpha - \varphi_\infty\|_{W_0^1(\mathbb{R}^2)} = O(\alpha^{-1/2})$,

(b) $\|\varphi_\alpha - \varphi_\infty\|_{L^2(\Omega)} = O(\alpha^{-3/2})$,

(c) $\|\varphi_\alpha - \varphi_\infty\|_{L^2(\partial\Omega)} = O(\alpha^{-1})$,

(d) $\|\vec{\nabla}\varphi_\alpha\|_{L^2(\partial\Omega)} = O(1)$.

Proof. Set $\eta = \varphi_\alpha - \varphi_\infty$. We have

$$-\Delta\eta - 2i\alpha^2\eta = 0 \quad \text{in } \Omega, \qquad -\Delta\eta = 0 \quad \text{in } \bar{\Omega}^c. \tag{5.1}$$

By Proposition 5.1 and Lemma 5.1, $\eta \in W_0^1(\mathbb{R}^2)$ but its normal derivative on $\partial\Omega$ has a jump that is equal to $d\varphi_\infty/dn$. We multiply both equations of (5.1) by $\bar{\eta}$, integrate, respectively, on Ω and $\bar{\Omega}^c$, use Green's formula, and add the resulting relations; we get

$$\int_{\mathbb{R}^2} |\vec{\nabla}\eta|^2 - 2i\alpha^2 \int_\Omega |\eta|^2 = -\int_{\partial\Omega} \frac{d\varphi_\infty}{dn}\,\bar{\eta}. \tag{5.2}$$

Let us take the modulus in (5.2). The left member is bounded from below by

$$(\|\vec{\nabla}\eta\|^2_{L^2(\mathbb{R}^2)} + 2\alpha^2\|\eta\|^2_{L^2(\Omega)})/2 \geqq (\|\eta\|^2_{W_0^1(\mathbb{R}^2)} + \|\alpha\eta\|^2_{L^2(\Omega)})/2$$

if $\alpha \geqq 1$ and if we set $\Lambda = \mathbb{R}^2$, $\Phi = \Omega$ in (2.2). As far as the right member is concerned, by Propositions 2.3 and Lemma 5.1 we can write

$$\left|\int_{\partial\Omega} \frac{d\varphi_\infty}{dn}\,\bar{\eta}\right|^2 \leqq \left\|\frac{d\varphi_\infty}{dn}\right\|^2_{L^2(\partial\Omega)} \|\eta\|^2_{L^2(\partial\Omega)} \leqq c\|\eta\|_{H^1(\Omega)}\|\eta\|_{L^2(\Omega)}$$

$$\leqq \frac{c}{\alpha}(\|\eta\|^2_{W_0^1(\mathbb{R}^2)} + \|\alpha\eta\|^2_{L^2(\Omega)}),$$

where c is a generic constant independent of α; parts (a) and (b) follow immediately. By Proposition 2.3, part (c) is a consequence of parts (a) and (b). By (5.1) and part (b) we obtain that $\|\Delta\eta\|_{L^2(\Omega)} = O(\alpha^{1/2})$, and consequently $\|\Delta\varphi_\alpha\|_{L^2(\mathbb{R}^2)} = O(\alpha^{1/2})$; since (part (a)) $\|\varphi_\alpha\|_{W_0^1(\mathbb{R}^2)} = O(1)$, we conclude by standard elliptic regularity results that $\|\varphi_\alpha\|_{H^2(\Omega)} = O(\alpha^{1/2})$. Consider a particular Ω_k and set $\zeta = \varphi_\alpha + C_k$; from the above results we have $\|\zeta\|_{H^2(\Omega_k)} = O(\alpha^{1/2})$ and $\|\zeta\|_{H^1(\Omega_k)} = O(\alpha^{-1/2})$; applying Proposition 2.3 to $\partial_l\zeta$, we obtain $\|\partial_l\zeta\|_{L^2(\partial\Omega)} = O(1)$, $l = 1, 2$; this proves part (d) of Proposition 5.2. □

Next we consider problem (4.5), (4.6). Let $\zeta \in C^\infty(\mathbb{R}^2)$ such that $\zeta(x) = 0$ for $|x| \leqq a$ and $\zeta(x) = 1$ for $|x| \geqq a+1$; a is a fixed number chosen in such a way that $\Omega \subset B(0, a-1)$. $g(x) = \Delta(\zeta(x)\ln|x|)$ is then a $C^\infty(\mathbb{R}^2)$ function, with bounded support and vanishing on Ω. By using variational methods as in the proofs of Proposition 5.1 and Lemma 5.1, we define uniquely w_α and $w_\infty \in W_0^1(\mathbb{R}^2)$ by the requirements

$$-\Delta w_\alpha = 2i\alpha^2\chi_\Omega w_\alpha + g \quad \text{in distribution in } \mathbb{R}^2, \tag{5.3}$$

$$w_\infty = 0 \quad \text{in } \Omega, \quad -\Delta w_\infty = g \quad \text{in distribution in } \bar{\Omega}^c. \tag{5.4}$$

By Propositions 2.1 and 2.2, there exist ξ_α and $\xi_\infty \in \mathbb{C}$ such that

$$w_\alpha(x) = \xi_\alpha + O(|x|^{-1}), \qquad w_\infty(x) = \xi_\infty + O(|x|^{-1}). \tag{5.5}$$

Now set

$$\varphi_\alpha = \zeta\ln|x| + w_\alpha - \xi_\alpha, \qquad \varphi_\infty = \zeta\ln|x| + w_\infty - \xi_\infty. \tag{5.6}$$

PROPOSITION 5.3. (a) *Problem* (4.5), (4.6) *possesses a unique solution* $(\varphi_\alpha, \xi_\alpha)$; φ_α *belongs to* $C^{1,\beta}(\mathbb{R}^2) \cap W^{2,p}_{\text{loc}}(\mathbb{R}^2)$, *for any* β, p *with* $0 < \beta < 1$, $1 \leqq p < +\infty$.

(b) *There exist a unique function* φ_∞ *and a unique constant* $\xi_\infty \in \mathbb{C}$ *satisfying the properties*: $\varphi_\infty \in H^1_{\mathrm{loc}}(\mathbb{R}^2)$, $\varphi_\infty = -\xi_\infty$ *on* Ω, φ_∞ *is harmonic on* $\bar{\Omega}^c$, $\varphi_\infty(x) = \ln|x| + O(|x|^{-1})$ *as* $|x| \to \infty$.

(c) $(\varphi_\alpha, \xi_\alpha)$ *and* $(\varphi_\infty, \xi_\infty)$ *are given by* (5.5), (5.6).

Proof. We only check part (a). Direct calculations show that $(\varphi_\alpha, \xi_\alpha)$, as defined by (5.5), (5.6), is the solution of problem (4.5), (4.6). Suppose it has two solutions and denote their difference by (v, η). Then, by Propositions 2.1 and 2.2, $v \in W^1_0(\mathbb{R}^2)$ and satisfies the equation $-\Delta v = 2i\alpha^2(v+\eta)$, which implies $v = -\eta$ on $\mathbb{R}^2$; since $v(x) = O(|x|^{-1})$ as $|x| \to \infty$, we get $\eta = 0$. Regularity results follow as in the proof of Proposition 3.1. □

We now study the convergence of φ_α toward φ_∞. The arguments of the proof of Proposition 5.2 apply to w_α and w_∞; in fact, as is easily checked, Proposition 5.2 is valid when we replace φ_α and φ_∞ by w_α and w_∞, respectively. By (5.6), it remains to estimate the term $(\xi_\alpha - \xi_\infty)$. To this aim, consider the domain $B(0, r)\backslash\bar{\Omega}$ for r large enough. By Green's formula, we obtain

$$\int_{\partial\Omega} \left\{ (w_\alpha - w_\infty) \frac{d\varphi_\infty}{dn} - \varphi_\infty \frac{d(w_\alpha - w_\infty)}{dn} \right\} = \int_{\partial B(0,r)} \left\{ (w_\alpha - w_\infty) \frac{d\varphi_\infty}{dn} - \varphi_\infty \frac{d(w_\alpha - w_\infty)}{dn} \right\}, \tag{5.7}$$

where the normal is exterior to Ω and $B(0, r)$. We use the asymptotic behaviours of φ_∞, w_α and w_∞ as $|x| \to \infty$. Standard arguments show that $\int_{\partial\Omega} \varphi_\infty \, d(w_\alpha - w_\infty)/dn = -\xi_\infty \int_{\partial\Omega} d(w_\alpha - w_\infty)/dn = 0$; furthermore, as $r \to \infty$, the right-hand side of (5.7) converges to $2\pi(\xi_\alpha - \xi_\infty)$. It follows from above that

$$|\xi_\alpha - \xi_\infty| = \left| \frac{1}{2\pi} \int_{\partial\Omega} (w_\alpha - w_\infty) \frac{d\varphi_\infty}{dn} \right| \leqq \frac{1}{2\pi} \left\| \frac{d\varphi_\infty}{dn} \right\|_{L^2(\partial\Omega)} \| w_\alpha - w_\infty \|_{L^2(\partial\Omega)} = O(\alpha^{-1}).$$

Summarizing this discussion, we have Proposition 5.4.

PROPOSITION 5.4. *Let* $(\varphi_\alpha, \xi_\alpha)$ *and* $(\varphi_\infty, \xi_\infty)$ *be given by Proposition* 5.3. *Then*:

(a) $\| \varphi_\alpha - \varphi_\infty \|_{W^1_0(\mathbb{R}^2)} = O(\alpha^{-1/2})$,

(b) $\| \varphi_\alpha - \varphi_\infty + (\xi_\alpha - \xi_\infty) \|_{L^2(\Omega)} = O(\alpha^{-3/2})$,

(c) $\| \varphi_\alpha - \varphi_\infty \|_{L^2(\partial\Omega)} = O(\alpha^{-1})$,

(d) $\| \vec{\nabla} \varphi_\alpha \|_{L^2(\partial\Omega)} = O(1)$,

(e) $|\xi_\alpha - \xi_\infty| = O(\alpha^{-1})$.

Remark 5.1. The estimates of Propositions 5.2 and 5.4 are identical except for part (b).

The next result, which is probably not optimal, will not be used later; we quote it without proof.

PROPOSITION 5.5. *Let* φ_α *and* φ_∞ *be defined either by Proposition* 5.2 *or Proposition* 5.3. *Then*

$$\| \varphi_\alpha - \varphi_\infty \|_{L^\infty(\mathbb{R}^2)} = O(\alpha^{-1/2}).$$

The following estimate, difficult and certainly not optimal, concerns the convergence of the normal derivative as $\alpha \to \infty$. We will prove it at the end of § 6.

PROPOSITION 5.6. *Let* φ_α *and* φ_∞ *be defined either by Proposition* 5.2 *or Proposition* 5.3. *Then* $\| \vec{\nabla}(\varphi_\alpha - \varphi_\infty) \|_{L^1(\partial\Omega)} = O((\alpha / \ln \alpha)^{-1/2})$.

We conclude this section by an example and a remark.

Example 5.1. The system consists of one conductor with circular section, i.e., $\Omega = B(0, R)$. For the nonphysical problem (4.2) we obtain the trivial solution $\varphi_\alpha = \varphi_\infty = -C_1$ on $\mathbb{R}^2$. More interesting is problem (4.5), (4.6) which we now consider. Here

$\varphi_\alpha(x) = \varphi_\infty(x) = \ln|x|$ for $|x| \geqq R$ and $\varphi_\infty(x) = -\xi_\infty = \ln R$ for $|x| \leqq R$, whereas, for $|x| \leqq R$, φ_α is of the form

$$\varphi_\alpha(x) = -\xi_\alpha + a_\alpha J_0((1+i)\alpha|x|). \tag{5.8}$$

In (5.8), J_0 denotes the Bessel function of order zero; the complex constants ξ_α and a_α are determined by imposing the continuity of φ_α and of its normal derivative on $\partial\Omega$. Explicit calculations using the asymptotic expansions of Bessel functions show that all estimates of Proposition 5.4 are optimal.

Remark 5.2. For $\Omega = B(0, R)$, let $(\varphi_\alpha, \xi_\alpha)$ be the solution of problem (4.5), (4.6) and $\tilde{\varphi}_\alpha$ be the solution of the initial problem (3.15), (3.16) with $C_1 = 1$. Clearly, $\tilde{\varphi}_\alpha = (1/\xi_\alpha)\varphi_\alpha$. By Proposition 5.4 and the above discussion of Example 5.1, $\lim_{\alpha\to\infty} \xi_\alpha = \xi_\infty = -\ln R$ and $\tilde{\varphi}_\alpha(x) = (1/\xi_\alpha)\ln|x|$ for $|x| \geqq R$. If $R = 1$, there is no possible convergence of $\tilde{\varphi}_\alpha$ as $\alpha \to \infty$ on any subset of $\bar{\Omega}^c$. In fact, for any Ω, there exists one and only one domain homothetic to Ω for which the same phenomenon occurs with $C_k = 1$, $k = 1, 2, \cdots, N$.

6. Some local estimates. In this section φ_α and φ_∞ will denote the functions defined either in Proposition 5.2 or in Proposition 5.3.

We first note that if $\Lambda \subset \mathbb{R}^2$ is a bounded open set such that $\bar{\Lambda}$ does not contain any corner of $\partial\Omega$, then for any m we have that φ_α and φ_∞ belong to $H^m(\Lambda \cap \Omega)$ and to $H^m(\Lambda \cap \bar{\Omega}^c)$. Loosely speaking, the restrictions of φ_α and φ_∞ to $\bar{\Omega}$ or to Ω^c are C^∞ except at the corners of $\partial\Omega$. These properties are consequences of standard regularity results for elliptic equations. To investigate properties in the neighborhood of $\partial\Omega$, we introduce the following local transformation of coordinates. Let $\Gamma \subset \partial\Omega$ be a smooth arc. One of its endpoints, say P, can be a corner of $\partial\Omega$. Γ admits the parametrization $(\gamma_1(t), \gamma_2(t))$ where t is the arclength parameter such that P corresponds to $t = 0$. With $\vec{n} = (n_1(t), n_2(t))$ denoting, as usual, the unit normal exterior to Ω, we set

$$(x_1, x_2) = (\gamma_1(\xi_2), \gamma_2(\xi_2)) + \xi_1(n_1(\xi_2), n_2(\xi_2)). \tag{6.1}$$

As is well known, (ξ_1, ξ_2) defines a local orthogonal curvilinear system of coordinates, the metric of which is given by the quadratic form

$$d\xi_1^2 + s^2(\xi_1, \xi_2)\, d\xi_2^2, \qquad s(\xi_1, \xi_2) = 1 + \frac{\xi_1}{R(\xi_2)}; \tag{6.2}$$

$R(\xi_2)$ is the radius of curvature of Γ, where $R(\xi_2) > 0$ if Ω is "convex" at the point of Γ with parameter $t = \xi_2$. We will denote by D_1^l and D_2^l the lth partial derivatives with respect to ξ_1 and ξ_2, respectively, so that, for example, the scalar product of two gradients is given by

$$\vec{\nabla} u \cdot \vec{\nabla} v = D_1 u \cdot D_1 v + D_2 u \cdot D_2 v / s^2. \tag{6.3}$$

To P, endpoint of Γ, or to any Q interior point of Γ, we associate an open rectangle T as shown in Figs. 6.1 and 6.2. In particular we suppose the following:

(a) The system of coordinates (ξ_1, ξ_2) is defined and regular in $\bar{T}$.

(b) $T \cap \Omega = \{x \in T \mid \xi_1 < 0\}$ and $T \cap \bar{\Omega}^c = \{x \in T \mid \xi_1 > 0\}$ are nonempty and satisfy the cone condition.

(c) $P \in \bar{T}$ or $Q \in T$.

LEMMA 6.1. *Let T be a rectangle as defined above and let $\theta : \mathbb{R}^2 \to \mathbb{R}_+$ be a function such that θ, $\partial_1\theta$, and $\partial_2\theta$ belong to $L^\infty(\mathbb{R}^2)$ and $\theta = 0$ in T^c. Then for any $m = 1, 2, 3, \cdots$, we have as α tends to infinity:*

(a) $\|\theta^m D_2^m(\varphi_\alpha - \varphi_\infty)\|_{W_0^1(\mathbb{R}^2)} = O(\alpha^{-1/2})$,

(b) $\|\theta^m D_2^m(\varphi_\alpha - \varphi_\infty)\|_{L^2(\Omega)} = O(\alpha^{-3/2})$,

(c) $\|\theta^m D_2^m(\varphi_\alpha - \varphi_\infty)\|_{L^2(\partial\Omega)} = O(\alpha^{-1})$.

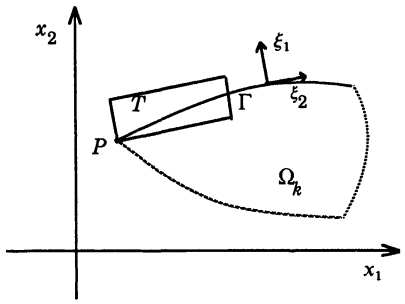


FIG. 6.1.

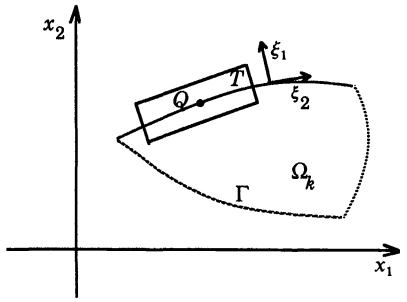


FIG. 6.2

Proof. We will prove Lemma 6.1 only for $m=1$ and for φ_α, φ_∞ defined by Proposition 5.2. The estimates relative to $m \geqq 2$ can be obtained with similar arguments by an induction process. The situation corresponding to Proposition 5.3 is treated easily by considering the function $\varphi_\alpha - \varphi_\infty + \xi_\alpha - \xi_\infty$; see Proposition 5.4(b). We start as in the proof of Proposition 5.2 with $\eta = \varphi_\alpha - \varphi_\infty$. Multiplying (5.1) by $v \in C_0^\infty(T)$, we obtain after integration by parts

$$\int_{\mathbb{R}^2} \vec{\nabla}\eta \cdot \vec{\nabla} v - 2i\alpha^2 \int_\Omega \eta v = -\int_{\partial\Omega} \frac{d\varphi_\infty}{dn} v. \tag{6.4}$$

In the coordinates (ξ_1, ξ_2), by using in particular (6.3), (6.4) becomes

$$\int_{\mathbb{R}^2} (sD_1\eta D_1 v + s^{-1} D_2 \eta D_2 v) - 2i\alpha^2 \int_{\tilde{\Omega}} s\eta v = -\int_{\mathbb{R}} D_1\varphi_\infty(0, \xi_2) v(0, \xi_2)\, d\xi_2 \tag{6.5}$$

where $\tilde{\Omega}$ is the image of $T \cap \Omega$ in the (ξ_1, ξ_2) plane. We set $v = -D_2 w$ in (6.5), where $w \in C_0^\infty(\tilde{T})$ and $\tilde{T}$ is the image of T in the (ξ_1, ξ_2) plane; after integration by parts, we obtain

$$\begin{aligned} &\int_{\mathbb{R}^2} (sD_1 D_2 \eta D_1 w + s^{-1} D_2^2 \eta D_2 w) - 2i\alpha^2 \int_{\tilde{\Omega}} sD_2 \eta w \\ &= -\int_{\mathbb{R}} D_1 D_2 \varphi_\infty(0, \xi_2) w(0, \xi_2)\, d\xi_2 \\ &\quad - \int_{\mathbb{R}^2} (D_2 s D_1 \eta D_1 w + D_2(s^{-1}) D_2 \eta D_2 w) + 2i\alpha^2 \int_{\tilde{\Omega}} D_2 s \eta w. \end{aligned} \tag{6.6}$$

In a first step, we assume that $\theta \in C_0^\infty(T)$, therefore we can set $w = \theta z$ with $z \in C_0^\infty(\tilde{T})$; after some elementary calculations, with $\Psi = D_2\eta = D_2(\varphi_\alpha - \varphi_\infty)$, we get

$$
\begin{aligned}
&\int_{\mathbb{R}^2} (sD_1(\theta\Psi)D_1 z + s^{-1}D_2(\theta\Psi)D_2 z) - 2i\alpha^2 \int_{\tilde{\Omega}} s\theta\Psi z \\
&\quad = -\int_{\mathbb{R}} \theta D_1 D_2\varphi_\infty(0, \xi_2)\, z(0, \xi_2)\, d\xi_2 \\
&\quad\quad - \int_{\mathbb{R}^2} (D_2 s D_1\eta D_1(\theta z) + D_2(s^{-1})D_2\eta D_2(\theta z)) + 2i\alpha^2 \int_{\tilde{\Omega}} D_2 s\eta\theta z \\
&\quad\quad + \int_{\mathbb{R}^2} \{sD_1\theta(\Psi D_1 z - zD_1\Psi) + s^{-1}D_2\theta(\Psi D_2 z - zD_2\Psi)\}.
\end{aligned}
\tag{6.7}
$$

From classical results of regularity $\theta D_1 D_2\varphi_\infty(0, \xi_2) \in C^\infty(\partial\Omega)$; therefore it follows from (6.7) that

$$
\mathscr{L}(\theta\Psi) \equiv -[D_1(sD_1) + D_2(s^{-1}D_2)](\theta\Psi) \in H^{-1}(\tilde{T});
$$

since $\theta\Psi$ has a compact support included in $\tilde{T}$, we deduce from the hypoellipticity of $\mathscr{L}$ that $\theta\Psi \in H_0^1(\tilde{T})$. Then by density we can set $z = \theta\bar{\Psi}$ in (6.7). In the last term of the right member, we replace $\theta\bar{\Psi}D\Psi$ by $\bar{\Psi}D(\theta\Psi) - \Psi\bar{\Psi}D\theta$; we use Proposition 2.3 and Proposition 5.2(a),(b) and obtain the estimate

$$
\begin{aligned}
&|\|\theta\Psi\|^2_{W_0^1(\mathbb{R}^2)} - 2i\alpha^2\|\theta\Psi\|^2_{L^2(\Omega)}| \\
&\leqq C\|\theta D_1 D_2\varphi_\infty\|_{L^2(\partial\Omega)}(\|\theta\Psi\|_{L^2(\Omega)}\|\theta\Psi\|_{H^1(\Omega)})^{1/2} + C\alpha^{-1/2}\|\theta^2\Psi\|_{W_0^1(\mathbb{R}^2)} \\
&\quad + C\alpha^{1/2}\|\theta^2\Psi\|_{L^2(\Omega)} + C\|\vec{\nabla}\theta\|_{L^\infty(\mathbb{R}^2)}(\alpha^{-1/2}\|\theta\Psi\|_{W_0^1(\mathbb{R}^2)} + \alpha^{-1}\|\vec{\nabla}\theta\|_{L^\infty(\mathbb{R}^2)})
\end{aligned}
\tag{6.8}
$$

where C is a generic constant independent of α and θ.

From results of Grisvard [3], $\delta D_1 D_2\varphi_\infty \in L^2(\partial\Omega)$, where $\delta(x)$ denotes the distance from x to the corners of $\partial\Omega$; clearly, $|\theta(x)| \leqq \delta(x)\|\vec{\nabla}\theta\|_{L^\infty(\mathbb{R}^2)}$ and $\|\theta D_1 D_2\varphi_\infty\|_{L^2(\partial\Omega)} \leqq C\|\vec{\nabla}\theta\|_{L^\infty(\mathbb{R}^2)}$. After a convenient treatment of the terms $\theta^2\Psi$ in (6.8) and by using classical inequalities, we obtain for C independent of α and θ

$$
\|\theta\vec{\nabla}\Psi\|_{L^2(\mathbb{R}^2)} \leqq C\alpha^{-1/2}\|\vec{\nabla}\theta\|_{L^\infty(\mathbb{R}^2)}, \qquad \|\theta\Psi\|_{L^2(\Omega)} \leqq C\alpha^{-3/2}\|\vec{\nabla}\theta\|_{L^\infty(\mathbb{R}^2)};
\tag{6.9}
$$

this proves parts (a) and (b) of Lemma 6.1 when $\theta \in C_0^\infty(T)$; part (c) follows from Proposition 2.3.

Now we consider the case where $\theta(x) = g(x) =$ distance from x to ∂T if $x \in T$ and $\theta(x) = 0$ if $x \in T^c$; then we can find a sequence of $\theta_n \in C_0^\infty(T)$ such that

$$
\theta_n \text{ converges to } g \text{ everywhere, and } \|\vec{\nabla}\theta_n\|_{L^\infty(\mathbb{R}^2)} \leqq \left(1 + \frac{1}{n}\right)\|\vec{\nabla}g\|_{L^\infty(\mathbb{R}^2)}.
$$

We obtain (6.9) for $\theta = g$ as the limit for $\theta = \theta_n$ by using the Lebesgue Theorem. The general situation is treated easily by noting that $|\theta(x)| \leqq g(x)\|\vec{\nabla}\theta\|_{L^\infty(\mathbb{R}^2)}$. □

LEMMA 6.2. *Let $\Lambda \subset \bar{\Omega}^c$ be an open bounded set. Then*

$$
\|\varphi_\alpha - \varphi_\infty\|_{L^2(\Lambda)} = O(\alpha^{-1}).
$$

Proof. Set $\eta = \varphi_\alpha - \varphi_\infty$ and let $w \in W_0^1(\bar{\Omega}^c)$ satisfy the relations

$$
\Delta w = \eta\chi_\Lambda \quad \text{in } \bar{\Omega}^c, \qquad w = 0 \quad \text{on } \partial\Omega,
\tag{6.10}
$$

where χ_Λ is the characteristic function of Λ. By Grisvard's regularity results [3], there exists a constant c, independent of η, such that $\|dw/dn\|_{L^2(\partial\Omega)} \leqq c\|\eta\|_{L^2(\Lambda)}$. Let r_0 be such that $\Omega \cup \Lambda \subset B(0, r_0)$. By Green's formula, we have for $r > r_0$

$$\|\eta\|^2_{L^2(\Lambda)} = \int_{\Omega^c \cap B(0,r)} \eta \Delta w = -\int_{\partial\Omega} \eta \frac{dw}{dn} + \int_{\partial B(0,r)} \left(\eta \frac{dw}{dn} - w\frac{d\eta}{dn}\right). \tag{6.11}$$

η and w belong to $W^1_0(\mathbb{R}^2)$ and are harmonic in $(\bar{B}(0, r_0))^c$. By Propositions 2.1 and 2.2, these functions and their normal derivatives behave, respectively, as do $O(1)$ and $O(|x|^{-2})$ as $|x|$ tends to infinity; we conclude that the second term of the right-hand side of (6.11) vanishes. Lemma 6.2 then follows from (6.11) and from Proposition 5.2(c) or Proposition 5.4(c). □

PROPOSITION 6.1. *Let $\Lambda \subset \bar{\Omega}^c$ be an open bounded set such that $\bar{\Lambda}$ contains no corner of $\partial\Omega$. Then for any $m = 0, 1, 2, \cdots$, we have*

$$\|\varphi_\alpha - \varphi_\infty\|_{H^m(\Lambda)} = O(\alpha^{-1}).$$

Proof. We remark that by Lemma 6.1(c) we have the estimate $\|\varphi_\alpha - \varphi_\infty\|_{H^m(\Gamma)} = O(\alpha^{-1})$ for any $m = 0, 1, 2, \cdots$ and any $\Gamma \subset \partial\Omega$ such that $\bar{\Gamma}$ contains no corner of $\partial\Omega$. Together with standard regularity results and Lemma 6.2, this implies Proposition 6.1. □

As an immediate consequence of the important Proposition 6.1, we have, for example, the following corollary.

COROLLARY 6.1. *For $\Gamma \subset \partial\Omega$ such that $\bar{\Gamma}$ contains no corner of $\partial\Omega$ and for any $m = 0, 1, 2, \cdots$, we have*

$$\|\vec{\nabla}(\varphi_\alpha - \varphi_\infty)\|_{H^m(\Gamma)} = O(\alpha^{-1}).$$

We conclude this section with the proof of Proposition 5.6. By Corollary 6.1, it suffices to prove Proposition 5.6 when we replace $\partial\Omega$ by a small arc Γ one endpoint of which is a corner; we can assume that $\Gamma \subset T \cap \partial\Omega$, where T is a rectangle as shown in Fig. 6.1. We use Proposition 5.2(a), Proposition 5.4(a), and Lemma 6.1(a) with $m = 1$ and $\theta = g$, where $g(x) = 0$ if $x \in T^c$ and $g(x)$ is the distance from x to ∂T if $x \in T$. We immediately deduce the relations $\|gD_1D_2\eta\|_{L^2(T)} = O(\alpha^{-1/2})$ and $\|gD_2^2\eta\|_{L^2(T)} = O(\alpha^{-1/2})$; from this last estimate and the fact that η is harmonic in $\bar{\Omega}^c$, we obtain that $\|gD_1^2\eta\|_{L^2(T\cap\bar{\Omega}^c)} = O(\alpha^{-1/2})$; then we can conclude that

$$\|g\partial_l\eta\|_{H^1(T\cap\bar{\Omega}^c)} = O(\alpha^{-1/2}), \qquad l = 1, 2. \tag{6.12}$$

Setting $w = \partial_l\eta$ and $\Lambda = T \cap \bar{\Omega}^c$, we have by a classical imbedding theorem and Schwarz's inequality

$$\begin{aligned} \|g|w|^2\|_{L^1(\partial\Lambda)} &\leqq C\left\{\|g|w|^2\|_{L^1(\Lambda)} + \sum_{j=1}^{2} \|\partial_j(g|w|^2)\|_{L^1(\Lambda)}\right\} \\ &\leqq C\left\{\|g\|_{L^\infty(\mathbb{R}^2)}\|w\|^2_{L^2(\Lambda)}\right. \\ &\qquad \left. + \sum_{j=1}^{2} (2\|\partial_j(gw)\|_{L^2(\Lambda)}\|w\|_{L^2(\Lambda)} + \|\partial_j g\|_{L^\infty(\Lambda)}\|w\|^2_{L^2(\Lambda)})\right\}. \end{aligned} \tag{6.13}$$

Since $\|w\|_{L^2(\Lambda)} = O(\alpha^{-1/2})$, it follows by (6.12) that the right-hand side of (6.13) is $O(\alpha^{-1})$; furthermore, we observe that along Γ, the arclength parameter ξ_2 is $O(g)$; consequently, we have obtained the estimate

$$\|\xi_2^{1/2}w\|_{L^2(\Gamma)} = O(\alpha^{-1/2}). \tag{6.14}$$

For $0<\gamma<1$, we can write $|w| = \xi_2^{(\gamma-1)/2}[\xi_2^{1/2}|w|]^{1-\gamma}|w|^{\gamma}$ so that by Hölder's inequality we obtain

$$\|w\|_{L^1(\Gamma)} \leqq \|\xi_2^{(\gamma-1)/2}\|_{L^2(\Gamma)}(\|\xi_2^{1/2}w\|_{L^2(\Gamma)})^{1-\gamma}\|w\|_{L^2(\Gamma)}^{\gamma}. \tag{6.15}$$

By Propositions 5.2(d) and 5.4(d), we know that $\|w\|_{L^2(\Gamma)}$ is uniformly bounded with respect to α; from (6.15) with $\gamma = 1/\ln\alpha$, we readily obtain the final estimate:

$$\|\partial_l(\varphi_\alpha - \varphi_\infty)\|_{L^1(\Gamma)} = \|w\|_{L^1(\Gamma)} = O((\alpha/\ln\alpha)^{-1/2}), \qquad l=1,2. \qquad \square$$

7. A boundary layer approximation. As in the previous section, φ_α will denote here the function defined either in Proposition 5.2 or Proposition 5.3. Our purpose is to study the behaviour of φ_α for large α in a conductor section Ω_k. We fix k, $1 \leqq k \leqq N$ and set $\Lambda = \Omega_k$. Furthermore, we set $\eta_\alpha = \eta = \varphi_\alpha + C_k$ for problem (4.2) and $\eta_\alpha = \eta = \varphi_\alpha + \xi$ for problem (4.5), (4.6).

We first remark that η_α satisfies in Λ the following Helmholtz homogeneous equation:

$$\Delta\eta_\alpha + 2i\alpha^2\eta_\alpha = 0 \quad \text{in } \Lambda. \tag{7.1}$$

It is well known that $F(x,y) = (1/4i)H_0^1((1+i)\alpha|x-y|)$ is a fundamental solution of this Helmholtz equation where H_0^1 is a Hankel function. It follows that η_α admits for $x \in \Lambda$ the representation

$$\eta_\alpha(x) = \int_{\partial\Lambda}\left\{\eta_\alpha(y)\frac{d}{dn_y}F(x,y) - F(x,y)\frac{d\eta_\alpha(y)}{dn_y}\right\}d\mu(y). \tag{7.2}$$

Then, by Propositions 5.2, 5.4, and by the asymptotic expansion formulae of the Hankel functions (see, for example, [5]), we immediately deduce the following proposition.

PROPOSITION 7.1. *Let $x \in \Lambda$ and d be the distance of x to $\partial\Lambda$. There exists a constant C_{mn} such that for $\alpha \geqq 1/d$ we have*

$$|\partial_1^m\partial_2^n\eta_\alpha(x)| \leqq C_{mn}\alpha^{m+n}\frac{e^{-\alpha d}}{\sqrt{\alpha d}}, \qquad m, n = 0, 1, 2, \cdots.$$

Now consider a smooth arc $\Gamma \subset \partial\Lambda$ such that the endpoints of Γ are not corners of $\partial\Lambda$. With the system of coordinates (ξ_1, ξ_2) introduced by (6.1), Γ is the set of points $(0, \xi_2)$, $0 \leqq \xi_2 \leqq \xi_{02}$. Furthermore, there exists $\delta > 0$ such that $V = \{x \mid -\delta < \xi_1 < 0,\ 0 < \xi_2 < \xi_{02}\} \subset \Lambda$ and such that the distance of the points of ∂V corresponding to $\xi_1 = -\delta$ to $\partial\Lambda$ is equal to δ. We define in $\bar{V}$ an approximate u_α of η_α by the relation

$$u_\alpha(\xi_1, \xi_2) = \frac{1}{\sqrt{s(\xi_1, \xi_2)}}e^{(1-i)\alpha\xi_1}\eta_\alpha(0, \xi_2), \tag{7.3}$$

where s is given by (6.2). We remark that $u_\alpha = \eta_\alpha$ on Γ.

As in § 6, D_1^l and D_2^l denote the partial derivatives of order l with respect to ξ_1 and ξ_2, whereas $R = R(\xi_2)$ is the radius of curvature of Γ.

PROPOSITION 7.2. *For any $m = 0, 1, 2, \cdots$ we have*

(a) $\|D_2^m(\eta_\alpha - u_\alpha)\|_{L^\infty(\bar{V})} = O(\alpha^{-3})$,

(b) $\|D_2^m D_1(\eta_\alpha - u_\alpha)\|_{L^\infty(\bar{V})} = O(\alpha^{-2})$.

Proof. We consider only the case $m = 0$, since the treatment of the general situation needs essentially the same tools. We use the expression of the Laplace operator in the (ξ_1, ξ_2) system of coordinates and write (7.1) in $\tilde{\Lambda}$, image of Λ in the (ξ_1, ξ_2) plane. Setting $w = \eta_\alpha - u_\alpha$, after some calculations we obtain

$$D_1(sD_1w) + 2i\alpha^2 sw = z, \tag{7.4}$$

$$z(\xi_1, \xi_2) = -D_2\left(\frac{1}{s}D_2\eta\right)(\xi_1, \xi_2) - \eta(0, \xi_2)\frac{e^{(1-i)\alpha\xi_1}}{4s^{3/2}(\xi_1, \xi_2)R^2(\xi_2)}. \tag{7.5}$$

From Lemma 6.1(b) and Proposition 6.1, we easily deduce that we have, uniformly with respect to $0 \leqq \xi_2 \leqq \xi_{02}$,

$$\|D_2^l \eta(\cdot, \xi_2)\|_{L^2(-\delta,0)} = O(\alpha^{-3/2}), \quad l = 1, 2, \quad |\eta(0, \xi_2)| = O(\alpha^{-1}). \tag{7.6}$$

Then (7.5) and (7.6) imply that

$$\|z(\cdot, \xi_2)\|_{L^2(-\delta,0)} = O(\alpha^{-3/2}) \quad \text{uniformly for } 0 \leqq \xi_2 \leqq \xi_{02}. \tag{7.7}$$

For ξ_2 fixed, we multiply (7.4) by $\bar{w}$; after an integration by parts, since $w(0, \xi_2) = 0$, we obtain

$$\left| -\int_{-\delta}^{0} s|D_1 w|^2 \, d\xi_1 + 2i\alpha^2 \int_{-\delta}^{0} s|w|^2 \, d\xi_1 \right| \leqq \|z(\cdot, \xi_2)\|_{L^2(-\delta,0)} \|w(\cdot, \xi_2)\|_{L^2(-\delta,0)} + |sw D_1 w(-\delta, \xi_2)|; \tag{7.8}$$

by (7.3), (7.6), and Proposition 7.1, the second term of the right-hand member of (7.8) is $O(e^{-\alpha\delta})$. Then (7.7) and (7.8) imply the estimate $\|w\|_{L^2(-\delta,0)} = O(\alpha^{-7/2})$ and $\|D_1 w\|_{L^2(-\delta,0)} = O(\alpha^{-5/2})$; with (7.4) and (7.7) we have, furthermore, $\|D_1^2 w\|_{L^2(-\delta,0)} = O(\alpha^{-3/2})$. We conclude by standard arguments. □

Remark 7.1. Results similar to those of Proposition 7.1 can be derived by directly using the Helmholtz equation (7.1) instead of the integral representation (7.2). This method is more complicated but can be generalized to the situation where α is variable, i.e., the conductivity is not constant.

Remark 7.2. Suppose that in the definition (7.3) of u_α we replace $s(\xi_1, \xi_2)$ by 1; then in Proposition 7.2 we lose one order, i.e., $O(\alpha^{-3})$ and $O(\alpha^{-2})$ are replaced, respectively, by $O(\alpha^{-2})$ and $O(\alpha^{-1})$.

Remark 7.3. Because of the exponential decay of η_α and u_α in the boundary layer, L^2 estimates are better than those for L^∞. From the proof of Proposition 7.2 we obtain, for example,

$$\|\eta_\alpha - u_\alpha\|_{L^2(V)} = O(\alpha^{-7/2}), \qquad \|\eta_\alpha - u_\alpha\|_{H^1(V)} = O(\alpha^{-5/2}). \tag{7.9}$$

8. An approximation of φ_α in $\bar{\Omega}^c$ satisfying a Robin boundary condition. Our purpose is to define a "cheap" approximation of φ_α. For the sake of briefness, we will consider only problem (4.2); problem (4.5), (4.6) can be treated in a similar way. To insure the validity of this approximation we must introduce a severe restriction on the regularity of Ω; in fact, we will suppose that

(8.1) $\partial\Omega$ is of class C^∞.

Let φ_α be the solution of problem (4.2). Suppose we know φ_α or an approximation Ψ_α of φ_α on $\partial\Omega$. Then, because of hypothesis (8.1), the results of § 7 allow us to define an explicit and simple approximation of φ_α in Ω.

Consider Proposition 7.2(b) with $m = 1$ and $\bar{V}$ being replaced by $\bar{V} \cap \partial\Omega$. Since $\eta_\alpha = \varphi_\alpha + C_k$, by (7.3) we immediately obtain the following proposition.

PROPOSITION 8.1. *Let φ_α be the solution of problem* (4.2) *and assume hypothesis* (8.1). *Then*

$$\left\| \frac{d\varphi_\alpha}{dn} - z_\alpha(\varphi_\alpha + C_k) \right\|_{L^\infty(\partial\Omega_k)} = O(\alpha^{-2}), \qquad 1 \leqq k \leqq N,$$

where $z_\alpha = (1-i)\alpha - 1/(2R)$ and R is the radius of curvature of $\partial\Omega_k$. □

Proposition 8.1 shows that φ_α satisfies approximately a Robin boundary condition on $\partial\Omega$. This leads us to introduce the following exterior problem. Find $\Psi_\alpha \in W_0^1(\bar{\Omega}^c)$

such that

$$\Delta\Psi_\alpha = 0 \quad \text{in } \bar{\Omega}^c, \qquad \frac{d\Psi_\alpha}{dn} = z_\alpha(\Psi_\alpha + C_k) \quad \text{on } \partial\Omega_k, \quad 1 \leqq k \leqq N, \tag{8.2}$$

where the unit normal $\vec{n}$ on $\partial\Omega$ is exterior to Ω.

PROPOSITION 8.2. *Let φ_α be the solution of problem* (4.2) *and assume hypothesis* (8.1). *Then*

(a) *Problem* (8.2) *has one and only one solution,*

(b) $\|\varphi_\alpha - \Psi_\alpha\|_{W_0^1(\bar{\Omega}^c)} = O(\alpha^{-5/2})$,

(c) $\|\varphi_\alpha - \Psi_\alpha\|_{L^2(\partial\Omega)} = O(\alpha^{-3})$.

Proof. Problem (8.2) admits the following variational formulation:

$$\int_{\Omega^c} \vec{\nabla}\Psi_\alpha \cdot \vec{\nabla} v + \int_{\partial\Omega} z_\alpha \Psi_\alpha v = -\sum_{k=1}^{N} C_k \int_{\partial\Omega_k} z_\alpha v \quad \forall v \in W_0^1(\bar{\Omega}^c). \tag{8.3}$$

Point (a) follows from (8.3) as an application of the Lax-Milgram Lemma. Set $w_\alpha = \Psi_\alpha - \varphi_\alpha$. Since $\varphi_\alpha \in W_0^1(\bar{\Omega}^c)$ and is harmonic in $\bar{\Omega}^c$, from 8.3 we obtain

$$\int_{\Omega^c} \vec{\nabla} w_\alpha \cdot \vec{\nabla} v + \int_{\partial\Omega} z_\alpha w_\alpha v = \sum_{k=1}^{N} \int_{\partial\Omega_k} \left(\frac{d\varphi_\alpha}{dn} - z_\alpha(\varphi_\alpha + C_k) \right) v \quad \forall v \in W_0^1(\bar{\Omega}^c). \tag{8.4}$$

We set $v = \bar{w}_\alpha$ in (8.4) and recall that $z_\alpha = (1-i)\alpha - 1/(2R)$; we take the imaginary part of this relation and obtain the estimate (c) by Proposition 8.1; replacing (8.4) with $v = \bar{w}_\alpha$ we conclude that $\|\nabla w_\alpha\|_{L^2(\bar{\Omega}^c)} = O(\alpha^{-5/2})$; together with part (c), this proves part (b). □

Remark 8.1. In addition to those shown in Proposition 8.2, other estimates can be obtained. For example, with some extra calculations, we can prove that

$$\|\varphi_\alpha - \Psi_\alpha\|_{L^\infty(\partial\Omega)} = O(\alpha^{-3}). \tag{8.5}$$

Furthermore, it is possible to extend φ_α to $\mathbb{R}^2$ by the boundary layer approximation introduced in § 7 and produce estimates relative to Ω. We will not pursue this. On one hand, these estimates are direct corollaries of the preceding ones or can be obtained by using the same tools; on the other hand, due to the very restrictive hypothesis (8.1), they are of limited interest.

Remark 8.2. Numerical tests show that for many practical applications, Ψ_α gives a very satisfactory approximation of φ_α if $\partial\Omega$ is regular. If $\partial\Omega$ has corners, the definition (8.2) of Ψ_α is still meaningful; however, we get only very poor theoretical error estimates, which are confirmed by numerical experiments.

Remark 8.3. If $\partial\Omega$ has corners, Proposition 8.1 is still valid when we replace $L^2(\partial\Omega_k)$ by $L^\infty(\Gamma)$ where Γ is a closed part of $\partial\Omega_k$ without singularities. In a forthcoming paper, we shall present successful numerical computations obtained with a method that takes advantage of this fact.

REFERENCES

[1] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] A. BOSSAVIT, *Le concept de "rigidité" dans la méthode des sous-structures; application à l'électronique*, in Numerical Methods for Engineering II, GAMNI II, Absi, Glowinski, Lascaux, Weysseyre, eds., Dunod, Bordas, Paris, 1980.

[3] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[4] S. I. HARIHARAN AND E. STEPHAN, *A boundary element method for a two-dimensional interface problem in electromagnetics*, Numer. Math. 42 (1983), pp. 311-322.
[5] E. JAHNKE, F. EMDE, AND F. LOSCH, *Tables of Higher Functions*, B. G. Teubner, Stuttgart, 1960.
[6] J. C. NEDELEC, *Approximation des equations intégrales en mécanique et en physique*, Rapport Centre de Mathématiques Appliquées, Ecole Polytechnique, Paris, 1977.
[7] A. SOMMERFELD, *Electrodynamics*, Academic Press, New York, 1952.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 593-630, May 1990
© 1990 Society for Industrial and Applied Mathematics
003

CONJECTURE ON THE STRUCTURE OF SOLUTIONS OF THE RIEMANN PROBLEM FOR TWO-DIMENSIONAL GAS DYNAMICS SYSTEMS*

TONG ZHANG† AND YUXI ZHENG‡

Abstract. Two-dimensional flow of polytropic gas with initial data being constant in each quadrant is considered. Under the assumption that each jump in initial data outside of the origin projects exactly one planar wave of shocks, centered rarefaction waves, or slip planes, it is proved that only 16 combinations of initial data are reasonable. For each combination, a conjecture on the structure of the solution in the whole space $t>0$ is given.

Key words. Riemann problem, shock, rarefaction waves, slip plane, selfsimilar solution, rarefactive pseudostationary flow

AMS(MOS) subject classifications. primary 35L65, 35L67; secondary 65M99

1. Introduction. Using the experiences of researching the Riemann problem and interactions of waves for one-dimensional gas dynamics systems [1]-[7], two-dimensional scalar conservation laws [8], two-dimensional steady flows [9], [10], and the regular reflection and Mach reflection [12], we consider the Riemann problem of two-dimensional gas dynamic systems. Both isentropic and adiabatic flow are considered.

The isentropic flow is modeled by the following system:

$$\begin{aligned} &\rho_t+(\rho u)_x+(\rho v)_y=0,\\ &(\rho u)_t+(\rho u^2+p)_x+(\rho uv)_y=0, \quad p=A\rho^\gamma, \quad \gamma>1, \quad A>0,\\ &(\rho v)_t+(\rho uv)_x+(\rho v^2+p)_y=0 \end{aligned} \tag{1.1}$$

and adiabatic flow,

$$\begin{aligned} &\rho_t+(\rho u)_x+(\rho v)_y=0,\\ &(\rho u)_t+(\rho u^2+p)_x+(\rho uv)_y=0,\\ &(\rho v)_t+(\rho uv)_x+(\rho v^2+p)_y=0,\\ &\left(\rho\left(e+\frac{u^2+v^2}{2}\right)\right)_t+\left(\rho u\left(h+\frac{u^2+v^2}{2}\right)\right)_x+\left(\rho v\left(h+\frac{u^2+v^2}{2}\right)\right)_y=0,\\ &e=\frac{p}{(\gamma-1)\rho}, \qquad h=e+\frac{p}{\rho}, \end{aligned} \tag{1.1$'$}$$

where ρ, (u, v), p denote density, velocity and pressure, respectively, and γ and A are constants.

The Riemann problems are defined, respectively, as follows:

$$(\rho, p, u, v)|_{t=0}=(\rho_i, p_i, u_i, v_i)\equiv ①, \tag{1.2$'$}$$

or in the ith quadrant of (x, y) plane $(i=1, 2, 3, 4)$,

$$(\rho, p, u, v)|_{t=0}=(\rho_i, p_i, u_i, v_i)\equiv ①, \tag{1.2$'$}$$

where ① are constant states.

* Received by the editors February 15, 1989; accepted for publication April 21, 1989. This work was supported by Academia Sinica, the National Science Foundation, National Natural Science Foundation of China, and Deutsche Forschungsgemeinschaft.

† Institute of Mathematics, Academia Sinica, Beijing, People's Republic of China.

‡ Department of Mathematics, University of California, Berkeley, California 94720.

Let us consider the selfsimilar solutions $(\rho, u, v)=(\rho(\xi, \eta), u(\xi, \eta), v(\xi, \eta))$ or $(\rho, p, u, v)=(\rho(\xi, \eta), p(\xi, \eta), u(\xi, \eta), v(\xi, \eta))$ $(\xi=x/t, \eta=y/t)$, called pseudostationary flow; then (1.1) or (1.1)′ changes, respectively, to

$$\begin{aligned} &-\xi\rho_\xi+(\rho u)_\xi-\eta\rho_\eta+(\rho v)_\eta=0,\\ &-\xi(\rho u)_\xi+(\rho u^2+p)_\xi-\eta(\rho u)_\eta+(\rho uv)_\eta=0,\\ &-\xi(\rho v)_\xi+(\rho uv)_\xi-\eta(\rho v)_\eta+(\rho v^2+p)_\eta=0, \end{aligned} \tag{1.3}$$

or

$$\begin{aligned} &-\xi\rho_\xi+(\rho u)_\xi-\eta\rho_\eta+(\rho v)_\eta=0,\\ &-\xi(\rho u)_\xi+(\rho u^2+p)_\xi-\eta(\rho u)_\eta+(\rho uv)_\eta=0,\\ &-\xi(\rho v)_\xi+(\rho uv)_\xi-\eta(\rho v)_\eta+(\rho v^2+p)_\eta=0,\\ &-\xi\left(\rho\left(e+\frac{u^2+v^2}{2}\right)\right)_\xi+\left(\rho u\left(h+\frac{u^2+v^2}{2}\right)\right)_\xi-\eta\left(\rho\left(e+\frac{u^2+v^2}{2}\right)\right)_\eta\\ &\quad+\left(\rho v\left(h+\frac{u^2+v^2}{2}\right)\right)_\eta=0, \end{aligned} \tag{1.3}′$$

and the Riemann problems change to boundary value problems at the infinity, i.e.,

$$(\rho, u, v) \text{ or } (\rho, p, u, v)\to\begin{cases}①\\②\\③\\④\end{cases} \text{ as } \xi^2+\eta^2\to\infty\begin{cases}\xi>0, & \eta>0,\\ \xi<0, & \eta>0,\\ \xi<0, & \eta<0,\\ \xi>0, & \eta<0.\end{cases}$$

We seek the solution in the whole (ξ, η) plane.

It is easy to show that the pseudostationary flow is transonic and must be supersonic at the infinity for bounded solutions. Considering the infinity as a Cauchy support, we construct the solution from infinity. The data at infinity are four constant states with four jumps. Solving these four jumps in the neighborhood of infinity, we obtain four planar waves, parallel to the coordinate axis, each of which consists, generally speaking, of three planar waves: a backward rarefaction wave $\tilde{R}$ or shock $\tilde{S}$, a slip line J, and a forward rarefaction wave $\vec{R}$ or shock $\vec{S}$. For simplicity, we assume that each jump of the data can be connected by exactly one planar wave of $\vec{R}$, $\vec{S}$, or J. Thus, in the neighborhood of infinity, the solution consists of four planar waves of R, S, and J, besides the four constant states. The problem will be classified according to the different combinations of these four waves. In § 3 we prove that only 16 cases of combinations are needed to be considered; six cases do not involve J, and 10 do involve J. Then we discuss the problem case by case in §§ 4–6. After some analysis, calculation, and demonstration, we get a conjecture on the structure of the solution in the whole (ξ, η) plane for each case.

We extend the supersonic planar flows coming from infinity along the stream lines. They should stop at a boundary of supersonic flow if they do not interact; otherwise they stop at the point where they interact. For the latter case, we solve the interaction of planar waves up to the boundary of supersonic flow, which consists of sonic curves, shocks, and slip lines. The boundary bounds a bounded domain of subsonic flow. Many beautiful structures appear in the solutions. When we solve the interaction of planar rarefaction waves, we must seek the global continuous solution of a Goursat problem with sonic points as the ends of its support, a problem of

degenerate hyperbolic system. When we solve the interaction of shocks, we meet the problem of reflection of an oblique shock, and the regular reflection and Mach reflection appear. To determine the subsonic flow we need to deal with a problem of a degenerate elliptic system with free boundary. Slip lines are more fascinating. They should end with a spiral [13]-[15] as they enter into the subsonic flow. In summary, the two-dimensional Riemann problem gathers many interesting open problems. Maybe there is a long way to go to solve all of these problems using pure analytical methods, but it is possible to check them by computation first.

DEFINITION. The pseudostationary flow is called rarefactive if the density is nonincreasing along the stream lines in the flow.

In summary, our conjecture on the solutions can be described as follows. The solutions are piecewise smooth except for a spiral of slip lines, and they consist of constant states, continuous rarefaction waves, shocks, and slip lines. There are no compressive continuous waves. These properties are similar to those of the one-dimensional case.

However, shocks may have bifurcation points and may vanish somewhere continuously. There is a subsonic domain in the flow, the boundary of which consists of sonic curves, slip lines, and/or shocks. The stream lines must go into the subsonic domain and focus at a node if there is no slip line in the domain; otherwise they may end with a spiral. All these features are substantially different from those of the one-dimensional case.

In addition, we have two conjectures for general pseudostationary flow as follows. (1) The pseudostationary flow is continuous on the whole plane if and only if it is continuous and rarefactive in a neighborhood of infinity. (2) The pseudostationary flow is smooth (i.e., C^1) on the whole plane if and only if it is a constant state.

2. Preliminaries.

2.1. Characteristics and standard forms. After simple calculation, the systems (1.3) and (1.3)′ can, for smooth solutions, be reduced to

$$\begin{pmatrix} U & \rho & 0 \\ p'/\rho & U & 0 \\ 0 & 0 & U \end{pmatrix}\begin{pmatrix} \rho_\xi \\ U_\xi \\ V_\xi \end{pmatrix}+\begin{pmatrix} V & 0 & \rho \\ 0 & V & 0 \\ p'/\rho & 0 & V \end{pmatrix}\begin{pmatrix} \rho_\eta \\ U_\eta \\ V_\eta \end{pmatrix}+\begin{pmatrix} 2\rho \\ U \\ V \end{pmatrix}=0,\tag{2.1}$$

[1]

and

$$\begin{pmatrix} U & 0 & \rho & 0 \\ 0 & 1 & \rho U & 0 \\ 0 & 0 & 0 & \rho U \\ 0 & U & \gamma p & 0 \end{pmatrix}\begin{pmatrix} \rho_\xi \\ p_\xi \\ U_\xi \\ V_\xi \end{pmatrix}+\begin{pmatrix} V & 0 & 0 & \rho \\ 0 & 0 & \rho V & 0 \\ 0 & 1 & 0 & \rho V \\ 0 & V & 0 & \gamma p \end{pmatrix}\begin{pmatrix} \rho_\eta \\ p_\eta \\ U_\eta \\ V_\eta \end{pmatrix}+\begin{pmatrix} 2\rho \\ \rho U \\ \rho V \\ 2\gamma p \end{pmatrix}=0,\tag{2.1}'$$

where $(U, V)=(u-\xi, v-\eta)$, which is called pseudovelocity. Their characteristic equations are

$$(V-\lambda U)[(V-\lambda U)^2-c^2(1+\lambda^2)]=0,$$

$$(V-\lambda U)^2[(V-\lambda U)^2-c^2(1+\lambda^2)]=0,$$

where $c=\sqrt{p'(\rho)}$ (isentropic) or $\sqrt{\gamma p/\rho}$ (adiabatic), which is called sound speed. So,

[1] It is symmetric when $p=\rho^3/3$.

either

$$V-\lambda U=0, \quad \text{i.e. } \lambda=\lambda_0\equiv\frac{V}{U} \text{ (pseudoflow characteristic),}$$

or

$$(V-\lambda U)^2=c^2(1+\lambda^2), \tag{2.2}$$

i.e.,

$$\lambda=\lambda_\pm\equiv\frac{UV\pm\sqrt{c^2(U^2+V^2-c^2)}}{U^2-c^2} \quad \text{(pseudowave or } \lambda \text{ characteristics),} \tag{2.3}$$

$$=\frac{V^2-c^2}{VU\mp\sqrt{c^2(U^2+V^2-c^2)}} \tag{2.4}$$

which are

real and distinct if and only if $U^2+V^2>c^2$ (supersonic),
real and the same if and only if $U^2+V^2=c^2$ (sonic),
complex conjugate if and only if $U^2+V^2<c^2$ (subsonic).

Obviously, bounded solutions (ρ, u, v) must be pseudosupersonic in the neighborhood of infinity.

There are three families of characteristic lines defined by

$$\frac{d\eta}{d\xi}=\lambda_i(\rho, U, V) \qquad (i=0, +, -),$$

which are called flow or $\lambda_\pm$ characteristic lines, respectively, in the supersonic domain for a given solution $(\rho(\xi, \eta), u(\xi, \eta), v(\xi, \eta))$. The flow characteristic lines, which are called stream lines also, can be oriented so that they come from the infinity. The $\lambda_\pm$ characteristic lines can never be tangent to the stream lines.

There are only stream lines in the subsonic flow.

The more interesting thing is the sonic curve. The $\lambda_\pm$ characteristic lines are tangent to each other at and only at sonic points and are perpendicular to the stream line there. What is the relation between the sonic curve and the $\lambda_\pm$ characteristic lines? In which cases will they coincide, contact, be perpendicular, or only intersect?

Let us deduce system (2.1) and (2.1)′ to characteristic form. The left eigenvectors corresponding to λ_0, $\lambda_\pm$ are, respectively,

$$l_0\|(0, U, V), \qquad l_\pm\|(V-\lambda_\pm U, \lambda_\pm\rho, -\rho),$$

$$l'_{01}\|(c^2, 0, 0, -1), \quad l'_{02}\|(0, U, V, 0), \qquad l'_\pm\|(0, \lambda_\pm c^2, -c^2, V-\lambda_\pm U).$$

The characteristic forms are

$$(c^2/\rho, U, V)\left(\frac{\partial}{\partial\xi}+\lambda_0\frac{\partial}{\partial\eta}\right)\begin{pmatrix}\rho\\U\\V\end{pmatrix}+(U+\lambda_0 V)=0, \tag{2.5}$$

$$\left(\pm\frac{\sqrt{c^2(U^2+V^2-c^2)}}{\rho}, -V, U\right)\left(\frac{\partial}{\partial\xi}+\lambda_\pm\frac{\partial}{\partial\eta}\right)\begin{pmatrix}\rho\\U\\V\end{pmatrix}+(\lambda_\pm U-V)=0; \tag{2.6}$$

(2.5)′ $\left(\frac{\partial}{\partial\xi}+\lambda_0\frac{\partial}{\partial\eta}\right)(p\rho^{-\gamma})=0$ (i.e., the smooth wave must be isentropic along each stream line),

$$\frac{1}{\rho}\left(\frac{\partial}{\partial\xi}+\lambda_0\frac{\partial}{\partial\eta}\right)p+\left(\frac{\partial}{\partial\xi}+\lambda_0\frac{\partial}{\partial\eta}\right)\left(\frac{U^2+V^2}{2}\right)+(U+\lambda_0 V)=0,$$

(2.6)′ $$\left(\pm\frac{\sqrt{c^2(U^2+V^2-c^2)}}{\rho c^2},-V,U\right)\left(\frac{\partial}{\partial\xi}+\lambda_\pm\frac{\partial}{\partial\eta}\right)\begin{pmatrix}p\\U\\V\end{pmatrix}+(\lambda_\pm U-V)=0.$$

The first-order system (2.5)-(2.6) consists of a quasilinear equation (2.5) and a mixed-type system of quasilinear equations (2.6). System (2.5)′-(2.6)′ is a hyperbolic system (2.5)′ coupled with a mixed-type system (2.6)′. It is well known that systems (2.5)-(2.6) and (2.5)′-(2.6)′ are both linearly degenerate for the flow characteristic and convex (genuinely nonlinear) for wave characteristic families.

2.2. Discontinuities. It is well known that along the discontinuity line $\eta=\eta(\xi)$, the Rankine-Hugoniot condition should be true, i.e.,

(2.7)
$$\begin{aligned}&[\rho U]\,d\eta=[\rho V]\,d\xi,\\&[\rho U^2+p]\,d\eta=[\rho UV]\,d\xi,\\&[\rho UV]\,d\eta=[\rho V^2+p]\,d\xi;\end{aligned}$$

(2.7)′
$$\begin{aligned}&[\rho U]\,d\eta=[\rho V]\,d\xi,\\&[\rho U^2+p]\,d\eta=[\rho UV]\,d\xi,\\&[\rho UV]\,d\eta=[\rho V^2+p]\,d\xi,\\&\left[\rho U\left(h+\frac{U^2+V^2}{2}\right)\right]d\eta=\left[\rho V\left(h+\frac{U^2+V^2}{2}\right)\right]d\xi,\end{aligned}$$

where $[Q]\equiv Q-Q_0$, i.e., the jump of Q across the discontinuity. Solving (2.7), we obtain either the linear discontinuity

(2.8)
$$\begin{aligned}&\frac{d\eta}{d\xi}=\frac{V}{U}=\frac{V_0}{U_0}\equiv\sigma_0\\&[\rho]=0\end{aligned}\qquad\text{(slip line)}$$

or nonlinear discontinuity

(2.9) $$\frac{d\eta}{d\xi}=\frac{U_0V_0\pm\sqrt{\bar c^2(U_0^2+V_0^2-\bar c^2)}}{U_0^2-\bar c^2}=-\frac{U-U_0}{V-V_0}\equiv\sigma_\pm,$$

(2.10) $$\rho_0(U_0\sigma_\pm-V_0)(V-V_0)=p-p_0,$$

where $\bar c^2=(\rho/\rho_0)([p]/[\rho])$. For the case of adiabatic flow, the only difference is that (2.8) is replaced by $[p]=0$ and the following equation should be added to (2.9) and (2.10):

(2.11) $$\frac{p}{p_0}=\frac{(\gamma+1)\rho-(\gamma-1)\rho_0}{(\gamma+1)\rho_0-(\gamma-1)\rho}>0.$$

It is easy to show that the nonlinear discontinuity can never be tangent to the stream line or the $\lambda_\pm$ characteristic lines. We call the compressive nonlinear discontinuity a shock or a shock wave.

2.3. Planar elementary waves and their sonic curve and characteristic lines. Planar elementary waves $(\rho(\xi), u(\xi), v(\xi))$ or $(\rho(\xi), p(\xi), u(\xi), v(\xi))$ involve:

(i) Constant states: $(\rho, u, v) = \text{const.}$ or $(\rho, p, u, v) = \text{const.}$

(ii) Backward and forward rarefaction waves:

$$\tilde{\tilde{R}}(\xi): \begin{cases} \xi = \lambda_{\mp} \equiv u \mp c, \\ \dfrac{du}{d\rho} = \mp \dfrac{c}{\rho}, \\ dv = 0, \end{cases} \quad \text{or} \quad \begin{cases} \xi = \lambda_{\mp} \equiv u \mp c, \\ \dfrac{du}{d\rho} = \mp \dfrac{c}{\rho}, \\ dv = 0, \\ d(p\rho^{-\gamma}) = 0 \quad \text{(i.e., } \tilde{\tilde{R}}(\xi) \text{ must be isentropic).} \end{cases}$$

(iii) Backward and forward shock waves:

$$\tilde{\tilde{S}}(\xi): \begin{cases} \xi = \sigma_{\mp} = u_l \mp \sqrt{\dfrac{\rho_r [p]}{\rho_l [\rho]}}, \\ \dfrac{[u]}{[\rho]} = \mp \sqrt{\dfrac{1}{\rho_r \rho_l} \dfrac{[p]}{[\rho]}}^2, \\ [V] = 0, \\ \rho_l \lesseqgtr \rho_r \quad \text{(entropy condition),} \end{cases}$$

or

$$\tilde{\tilde{S}}(\xi): \begin{cases} \xi = \sigma_{\mp} = u_l \mp \sqrt{\dfrac{\rho_r}{\rho_l} \dfrac{[p]}{[\rho]}}, \\ \dfrac{[u]}{[\rho]} = \mp \sqrt{\dfrac{1}{\rho_r \rho_l} \dfrac{[p]}{[\rho]}}^2, \\ [v] = 0, \\ \dfrac{p_l}{p_r} = \dfrac{(\gamma+1)\rho_l - (\gamma-1)\rho_r}{(\gamma+1)\rho_r - (\gamma-1)\rho_l} > 0, \\ p_l \lesseqgtr p_r \Leftrightarrow \rho_l \lesseqgtr \rho_r \quad \text{(entropy condition).} \end{cases}$$

(iv) Slip lines:

$$J(\xi): \begin{cases} \xi = u_l = u_r, \\ [\rho] = 0, \end{cases} \quad \text{or} \quad \begin{cases} \xi = u_l = u_r, \\ [p] = 0. \end{cases}$$

Let us consider their sonic curve and characteristic lines.

(i) Constant state $(\rho, u, v) = (\rho_0, u_0, v_0)$ or $(\rho, p, u, v) = (\rho_0, p_0, u_0, v_0)$. Their sonic curve is a circle:

$$(\xi - u_0)^2 + (\eta - v_0)^2 = c_0^2.$$

The flow is subsonic inside the circle and supersonic outside the circle. The stream lines satisfy

$$\frac{d\eta}{d\xi} = \frac{\eta - v_0}{\xi - u_0}.$$

[2] This is equivalent to $u_l \mp \sqrt{(\rho_r/\rho_l)([p]/[\rho])} = u_r \mp \sqrt{(\rho_l/\rho_r)([p]/[\rho])}$.

Integrating it, we obtain

$$\frac{\eta - v_0}{\xi - u_0} = \text{const.},$$

which are all rays starting from infinity and focusing at the center of the sonic circle.

The wave characteristic lines satisfy

$$\frac{d\eta}{d\xi} = \lambda_{\pm},$$

where $\lambda_{\pm}$ satisfy

$$[(\eta - v_0) - \lambda_{\pm}(\xi - u_0)]^2 = c_0^2(1 + \lambda_{\pm}^2).$$

Differentiating them along their own characteristic lines, we obtain

$$-2[(\eta - v_0) - \lambda_{\pm}(\xi - u_0)](\xi - u_0)\frac{d\lambda_{\pm}}{d\xi} = 2\lambda_{\pm}c_0^2\frac{d\lambda_{\pm}}{d\xi};$$

i.e.,

$$\pm\sqrt{c_0^2[(\xi - u_0)^2 + (\eta - v_0)^2 - c_0^2]}\,\frac{d\lambda_{\pm}}{d\xi} = 0,$$

so

$$\frac{d\lambda_{\pm}}{d\xi} = 0,$$

which means the wave characteristic lines are straight lines. Because they are perpendicular to the stream lines at the sonic circle, they must be the tangent lines of the

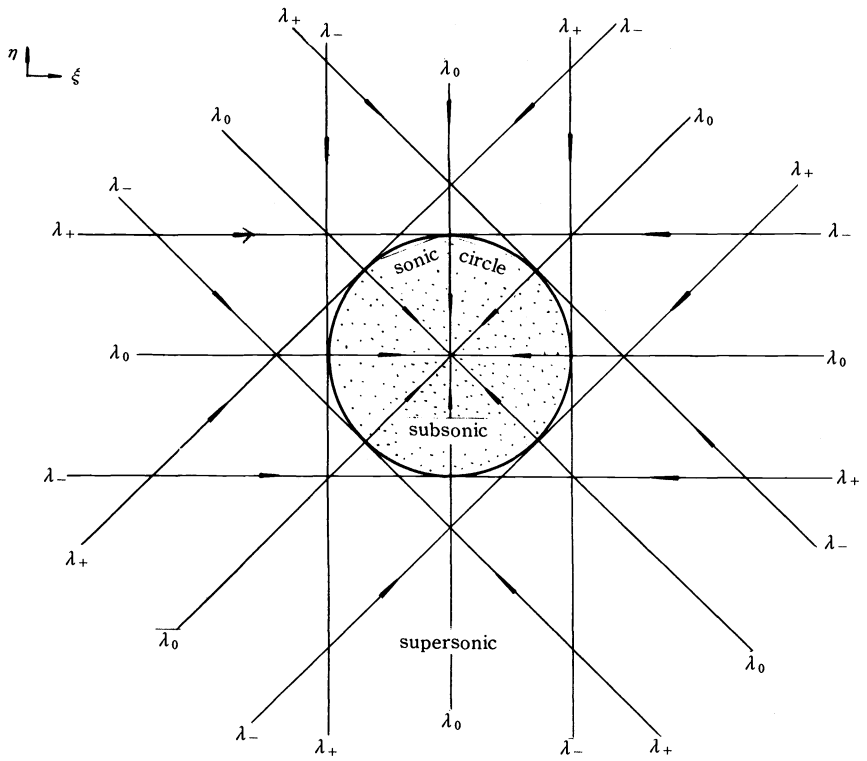


FIG. 2.1

circle. It is easy to establish that the clockwise rays correspond to λ_- and the counterclockwise rays correspond to λ_+. We can orient them so that they come from infinity and end at the sonic circle (Fig. 2.1).

(ii) $\vec{R}(\xi)_{2,1}$ connecting ② and ① (Figs. 2.2 and 2.3):

$$\vec{R}(\xi)_{2,1}:\begin{cases}\xi = u+\sqrt{p'(\rho)}, & (\xi_2 \equiv u_2+c_2 \leqq \xi \leqq u_1+c_1 \equiv \xi_1),\\ u = u_1+\displaystyle\int_{\rho_1}^{\rho}\frac{\sqrt{p'(\rho)}}{\rho}\,d\rho, & (0 \leqq \rho_2 \leqq \rho \leqq \rho_1),\\ v = v_1.\end{cases}$$

The sonic curve is a straight segment

$$\eta = v_1 \qquad (\xi_2 \leqq \xi \leqq \xi_1),$$

which we call the sonic stem.

The stream lines satisfy

$$\frac{d\eta}{d\xi} = \frac{\eta - v_1}{\sqrt{p'(\rho)}}$$

and we have

$$\frac{d\xi}{d\rho} = \frac{du}{d\rho}+\frac{1}{2}\frac{p''(\rho)}{\sqrt{p'(\rho)}} = \frac{\sqrt{p'}}{\rho}+\frac{1}{2}\frac{p''(\rho)}{\sqrt{p'(\rho)}} = \frac{\sqrt{A\gamma}}{2}(\gamma+1)\rho^{(\gamma-3)/2},$$

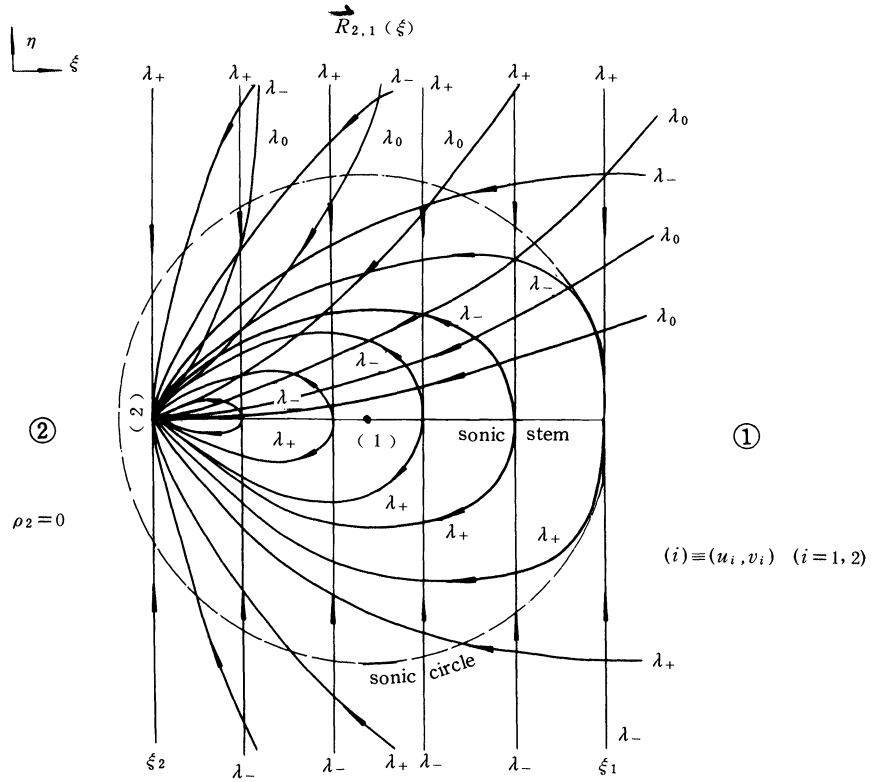


FIG. 2.2

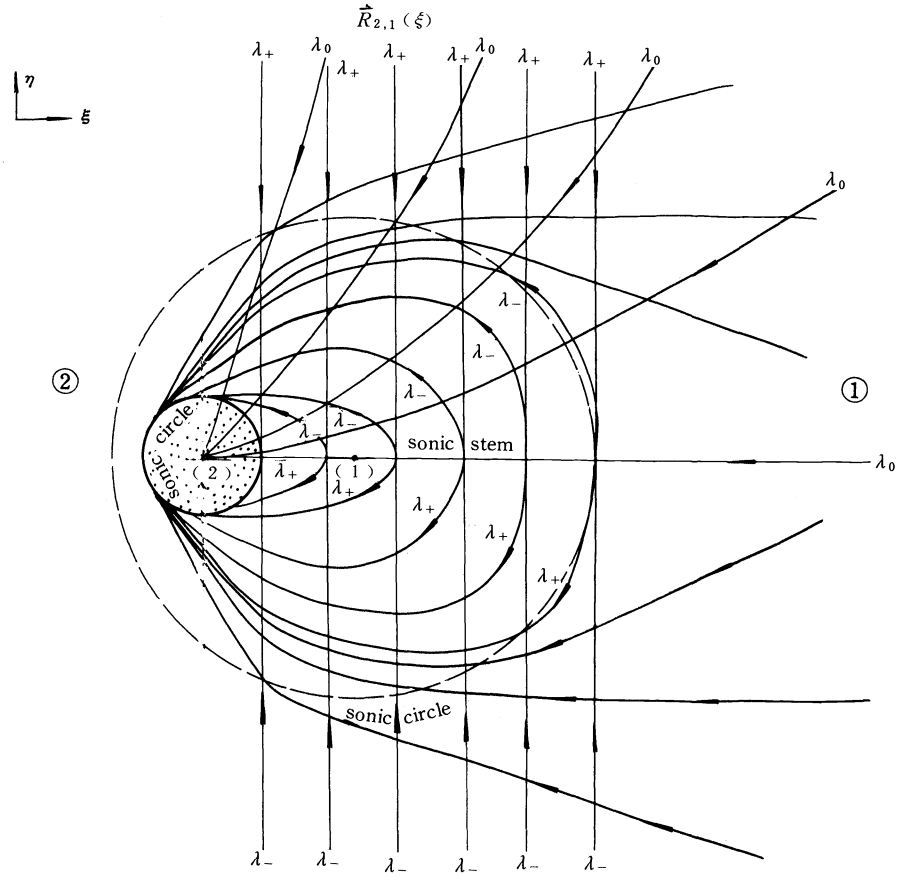


FIG. 2.3

so

$$\frac{d\eta}{d\rho}=\frac{(\gamma+1)(\eta-v_1)}{2\rho}.$$

Integrating them, we get

$$\xi-u_1=\sqrt{A\gamma}\rho_1^{(\gamma-1)/2}+\sqrt{A\gamma}\,\frac{\gamma+1}{\gamma-1}(\rho^{(\gamma-1)/2}-\rho_1^{(\gamma-1)/2}),$$

$$\eta-v_1=c\rho^{(\gamma+1)/2}(=c(\xi-u_2)^{(\gamma+1)/(\gamma-1)} \text{ if } \rho_2=0),$$

where c is an arbitrary constant. The sonic stem is located on the stream line $c=0$ and is perpendicular to the $\lambda_\pm$ characteristic lines there. Here the situation is totally different from that of constant states where the sonic curve is tangent to the $\lambda_\pm$ characteristic lines. The other stream lines are symmetric about it and all intersect at point (u_2, v_2). (Figure 2.2 is for $\rho_2=0$, Fig. 2.3 is for $\rho_2>0$. We use (i) to denote the point (u_i, v_i).)

The $\lambda_\pm$ characteristic lines satisfy

$$\frac{d\eta}{d\xi}=\lambda_\pm,$$

where

$$\lambda_+ = \infty, \quad \lambda_- = \frac{(\eta - v_1)^2 - p'}{2\sqrt{p'}(\eta - v_1)}, \quad (\eta > v_1),$$

$$\lambda_+ = \frac{(\eta - v_1)^2 - p'}{2\sqrt{p'}(\eta - v_1)}, \quad \lambda_- = \infty, \quad (\eta < v_1).$$

We need only to consider

$$\frac{d\eta}{d\xi} = \frac{(\eta - v_1)^2 - p'}{2\sqrt{p'}(\eta - v_1)}.$$

As $d\xi = ((2p' + p'')/(2\rho\sqrt{p'}))\, d\rho$, we have

$$\frac{d(\eta - v_1)^2}{d\rho} = \frac{\gamma + 1}{2\rho}[(\eta - v_1)^2 - \gamma\rho^{\gamma - 1}].$$

Integrating it, we get

$$(\eta - v_1)^2 = \rho^{(\gamma+1)/2}\left(c - \frac{\gamma(\gamma+1)}{\gamma - 3}\rho^{(\gamma-3)/2}\right) \qquad (\gamma \neq 3)$$

or

$$(\eta - v_1)^2 = \rho^2(c - \ln \rho) \qquad (\gamma = 3).$$

All of the $\lambda_\pm$ characteristic lines focus at (u_2, v_2) (see Fig. 2.2).

If ② is not a vacuum ($\rho_2 > 0$), then the sonic curves and characteristic lines of ②, $\vec{R}_{2,1}(\xi)$, and ① are as pictured in Fig. 2.3. Figure 2.3 can be made by combining Figs. 2.1 and 2.2. The sonic curve is made of the sonic circle of ② and the sonic stem of $\vec{R}_{2,1}(\xi)$. The sonic circle of ① is located in $\xi \leqq \xi_1$; therefore it is an imaginary one.

(iii) $\vec{S}_{2,1}(\xi)$ connecting ② and ① (Fig. 2.4):

$$\vec{S}_{2,1}(\xi): \begin{cases} \xi = \sigma_+ = u_2 + \sqrt{\dfrac{\rho_1}{\rho_2} p'_{12}} = u_1 + \sqrt{\dfrac{\rho_2}{\rho_1} p'_{12}}, \\ v_2 = v_1, \\ \rho_2 > \rho_1 \Leftrightarrow u_2 > u_1, \end{cases}$$

or

$$\vec{S}_{2,1}(\xi): \begin{cases} \xi = \sigma_+ = u_2 + \sqrt{\dfrac{\rho_1}{\rho_2} p'_{12}} = u_1 + \sqrt{\dfrac{\rho_2}{\rho_1} p'_{12}}, \\ v_2 = v_1, \\ \dfrac{p_2}{p_1} = \dfrac{(\gamma+1)\rho_2 - (\gamma-1)\rho_1}{(\gamma+1)\rho_1 - (\gamma-1)\rho_2} > 0, \\ p_2 > p_1 \Leftrightarrow \rho_2 > \rho_1 \Leftrightarrow u_2 > u_1, \end{cases}$$

where $p'_{12} \equiv (p_2 - p_1)/(\rho_2 - \rho_1)$.

Obviously, $u_2 + c_2 > \sigma_+ > u_1 + c_1$, $\sigma_+ > u_2$, which means that the sonic circle of ① is located in $\xi < \sigma_+$ (so, it is an imaginary one), and the sonic circle of ② is divided into two parts by the shock (so, it is partly imaginary). The characteristic lines of ① and ② are determined by these two sonic circles, respectively.

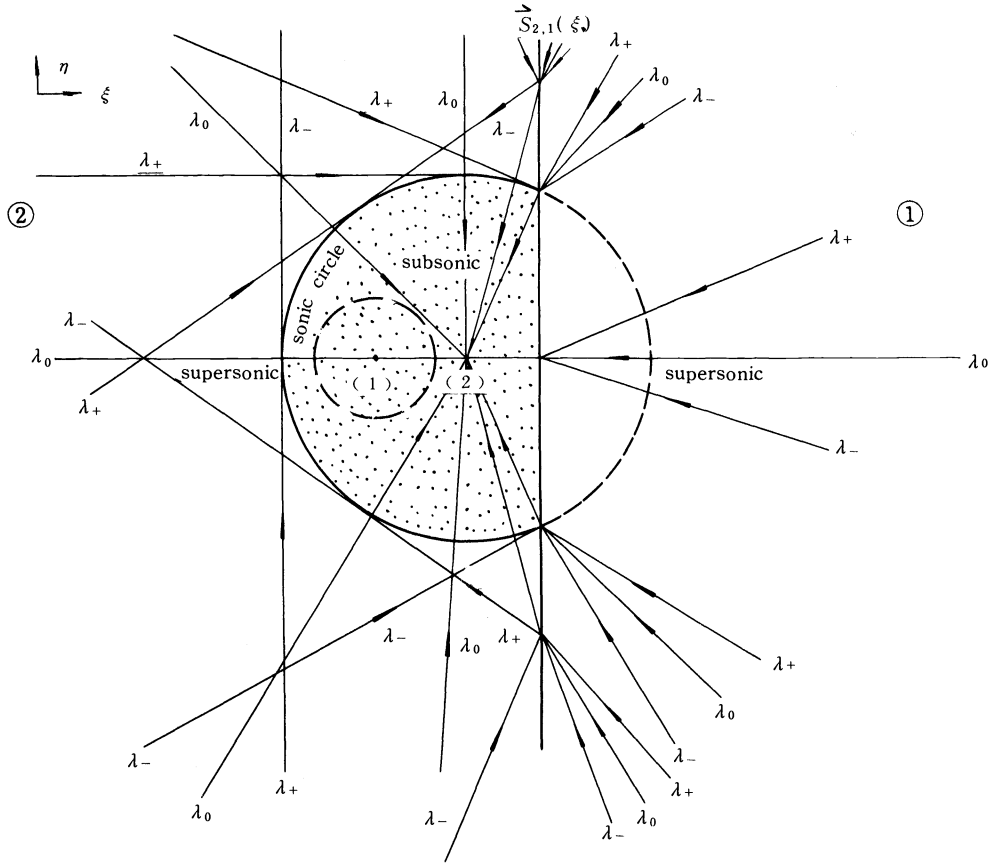


FIG. 2.4

(iv) $J_{2,1}(\xi)$ connecting ② and ① (Fig. 2.5):

$$J_{2,1}(\xi):\begin{cases}\xi = u_2 = u_1,\\ \rho_2 = \rho_1,\end{cases} \quad \text{or} \quad \begin{cases}\xi = u_2 = u_1,\\ p_2 = p_1.\end{cases}$$

It seems that we cut the constant states along the stream line $\xi = u_1 = u_2$ into two parts, and slip them up or down.

Completely analogously, we can deal with the remaining planar elementary waves along the other directions.

3. Classification. Under our assumptions, there are only four planar elementary waves besides the constant states in a neighborhood of infinity.

We now analyze which combinations of these four waves are possible.

We consider the combination where J does not appear first. Thus $\rho_i \neq \rho_j$ $(i \neq j,\ i, j = 1, 2, 3, 4)$. It is easy to show that there are only the following three combinations of ρ_i's:

$$\begin{array}{ccc} \rho_2 < \rho_1 & \rho_2 < \rho_1 & \rho_2 < \rho_1 \\ \vee \quad \vee & \wedge \quad \vee & \vee \quad \vee \\ \rho_3 < \rho_4 & \rho_3 > \rho_4 & \rho_2 > \rho_4 \end{array}$$

and that all other combinations of ρ_i's can be reduced to the three above by coordinate transformations.

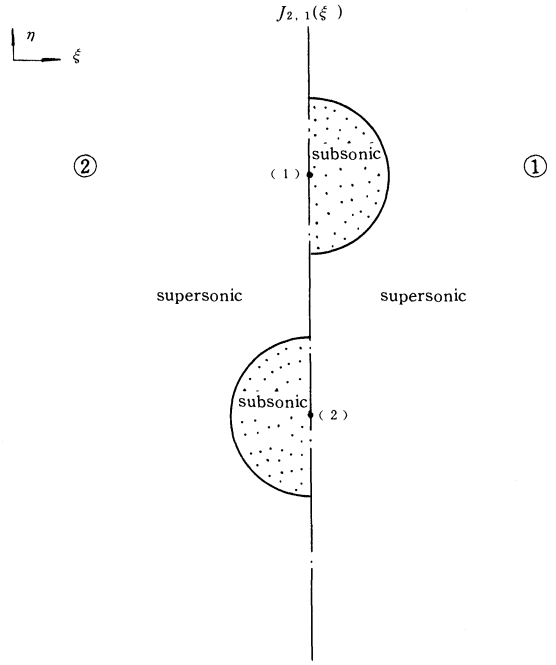


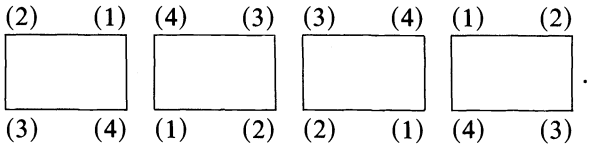
FIG. 2.5

For each of the above three possible combinations of ρ_i's, there are only two further possibilities:

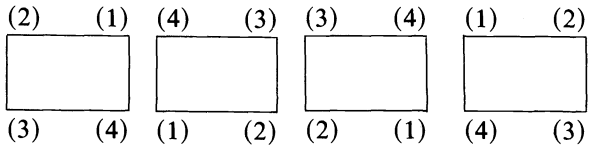
$$\rho_2 < \rho_1, \qquad u_2 < u_1: \vec{R}(\xi),$$

$$\rho_2 < \rho_1, \qquad u_2 > u_1: \tilde{S}(\xi).$$

It is easy to show that $(\xi, \eta) = (i) \equiv (u_i, v_i)$ $(i = 1, 2, 3, 4)$ must be located on the four corners of a rectangle, and there are only four possibilities as follows:



Combining all facts mentioned above, we have the following table:



$\rho_2 < \rho_1$	$\vec{R}$	$\tilde{S}$	$\vec{R}$	$\tilde{S}$
$\vee \quad \vee$	$\vec{R} \quad \vec{R}$	$\tilde{S} \quad \tilde{S}$	$\tilde{S} \quad \tilde{S}$	$\vec{R} \quad \vec{R}$
$\rho_3 < \rho_4$	$\vec{R}$	$\tilde{S}$	$\vec{R}$	$\tilde{S}$
$\rho_2 < \rho_1$	$\vec{R}$	$\tilde{S}$	$\vec{R}$	$\tilde{S}$
$\wedge \quad \vee$	$\tilde{R} \quad \vec{R}$	$\vec{S} \quad \tilde{S}$	$\vec{S} \quad \tilde{S}$	$\tilde{R} \quad \vec{R}$
$\rho_3 > \rho_4$	$\tilde{R}$	$\vec{S}$	$\tilde{R}$	$\vec{S}$

$\rho_2 < \rho_1$	$\vec{R}$	$\tilde{S}$	$\vec{R}$	$\tilde{S}$
$\vee \quad \vee$	$\vec{R} \quad \vec{R}$	$\tilde{S} \quad \tilde{S}$	$\tilde{S} \quad \tilde{S}$	$\vec{R} \quad \vec{R}$
$\rho_3 > \rho_4$	$\tilde{R}$	$\vec{S}$	$\tilde{R}$	$\vec{S}$

But the bottom set of combinations are impossible. In fact, taking as an example, $\vec{R}\overset{\vec{R}}{\tilde{R}}\vec{R}$, we have

$$v_1 = v_2, \quad v_2 - v_3 = \int_{\rho_3}^{\rho_2} \frac{\sqrt{p'}}{\rho}\, d\rho, \quad v_3 = v_4, \quad v_4 - v_1 = \int_{\rho_1}^{\rho_4} \frac{\sqrt{p'}}{\rho}\, d\rho;$$

thus

$$\int_{\rho_3}^{\rho_4} \frac{\sqrt{p'}}{\rho}\, d\rho = \int_{\rho_4}^{\rho_1} \frac{\sqrt{p'}}{\rho}\, d\rho,$$

which is obviously impossible.

For $\tilde{S}\overset{\tilde{S}}{\vec{S}}\tilde{S}$, we have

$$v_1 = v_2, \quad v_2 - v_3 = \sqrt{\frac{1}{\rho_2\rho_3} p'_{23}(\rho_2 - \rho_3)}, \quad v_3 = v_4, \quad v_4 - v_1 = \sqrt{\frac{1}{\rho_1\rho_4} p'_{14}(\rho_1 - \rho_4)};$$

thus

$$\sqrt{\frac{1}{\rho_2\rho_3} p'_{23}(\rho_2 - \rho_3)} = \sqrt{\frac{1}{\rho_1\rho_4} p'_{14}(\rho_1 - \rho_4)},$$

which is also easily shown to be impossible.

Analogously, the remaining cases are impossible. □

Obviously, the two columns to the right of the table are similar (i.e., coordinate rotations can transform one to the other); therefore there are only six cases that do not involve J: two for four R's, two for four S's, and two for two R's and two S's.

Now, we consider cases involving J.

(1) Four J's. There are only four subcases (Fig. 3.1). The two rows are similar in the figure. Therefore there are only two subcases.

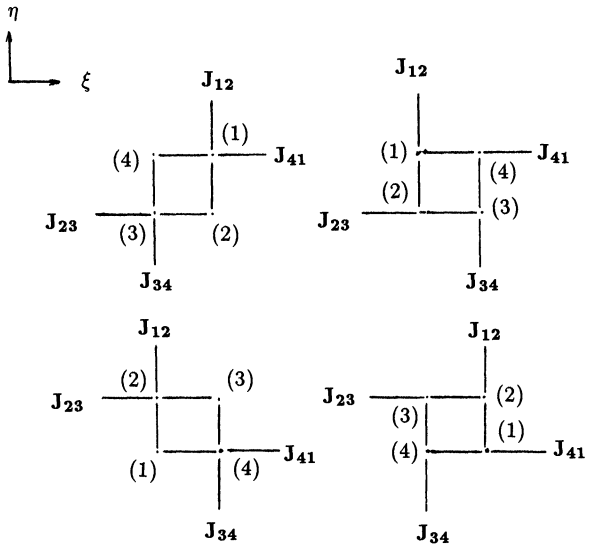


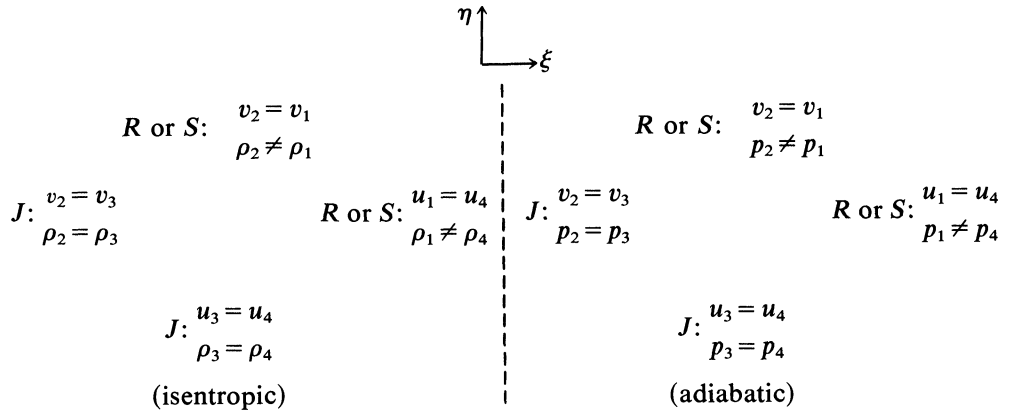
FIG. 3.1

(2) Three J's. From three J's we know that $\rho_1=\rho_2=\rho_3=\rho_4$ (isentropic) or $p_1=p_2=p_3=p_4$ (adiabatic), which contradicts the fourth wave.

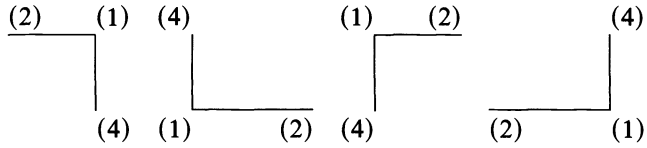
(3) Two J's.

(i) Two J's are neighbors. Without loss of generality, we assume that the two J's are in the directions of $-\xi$ and $-\eta$. We need to find out the wave types in the other two directions.

Observe the following table:



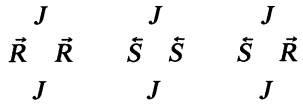
We have (1) = (3), and $\rho_2=\rho_3=\rho_4$ (isentropic) or $p_2=p_3=p_4$ (adiabatic). Thus



$\rho_2=\rho_4<\rho_1$	(isentropic)	$\vec{R}$	$\tilde{S}$	$\tilde{S}$	$\vec{R}$
or		$J\ \vec{R}$	$J\ \tilde{S}$	$J\ \vec{R}$	$J\ \tilde{S}$
$p_2=p_4<p_1$	(adiabatic)	J	J	J	J
$\rho_2=\rho_4>\rho_1$	(isentropic)	$\tilde{R}$	$\vec{S}$	$\vec{S}$	$\tilde{R}$
or		$J\ \tilde{R}$	$J\ \vec{S}$	$J\ \tilde{R}$	$J\ \vec{S}$
$p_2=p_4>p_1$	(adiabatic)	J	J	J	J

The two columns to the right are similar, so there are only six cases.

(ii) Two J's are not neighbors. It is easy to show that there are only three cases:

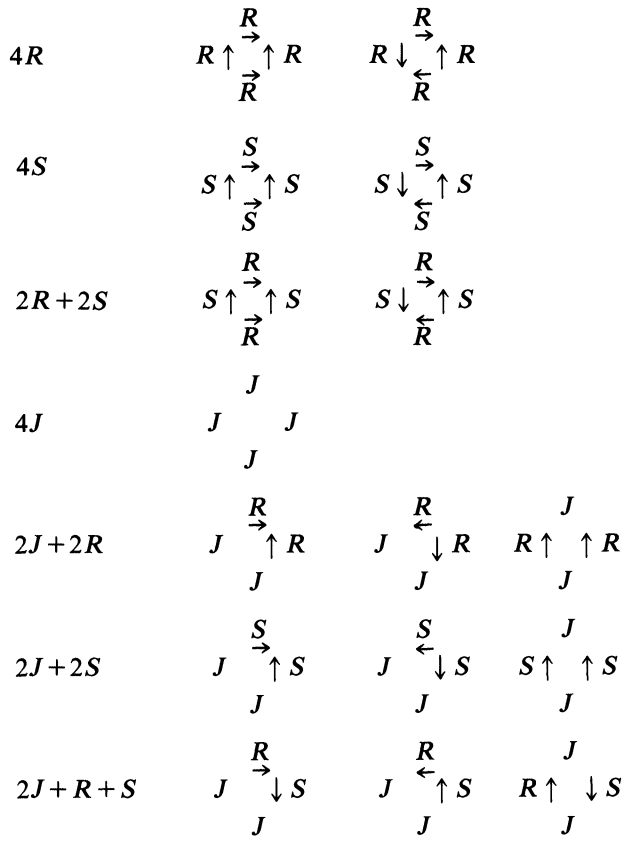


(4) One J:

$$J:\ u_2=u_1$$
$$R \text{ or } S:\ u_2=u_3 \qquad R \text{ or } S:\ u_1=u_4$$
$$R \text{ or } S:\ u_3\neq u_4.$$

It is obviously impossible.

In sum, we have 16 cases that need to be considered as follows:



We discuss them case by case in the forthcoming sections.

4. Four rarefaction waves. These must be isentropic. There are two cases.

(1) Four $\vec{R}$'s. The sonic circles and the sonic stems of $\vec{R}$'s are pictured in Fig. 4.1. The inequalities of ρ_i are

$$\rho_3 < \begin{matrix}\rho_2 \\ \rho_4\end{matrix} < \rho_1.$$

From

$$u_1 - u_2 = \int_{\rho_2}^{\rho_1} (\sqrt{p'}/\rho)\, d\rho, \qquad v_1 - v_2 = 0,$$

$$u_2 - u_3 = 0, \qquad v_2 - v_3 = \int_{\rho_3}^{\rho_2} (\sqrt{p'}/\rho)\, d\rho,$$

$$u_3 - u_4 = \int_{\rho_4}^{\rho_3} (\sqrt{p'}/\rho)\, d\rho, \qquad v_3 - v_4 = 0,$$

$$u_4 - u_1 = 0, \qquad v_4 - v_1 = \int_{\rho_1}^{\rho_4} (\sqrt{p'}/\rho)\, d\rho,$$

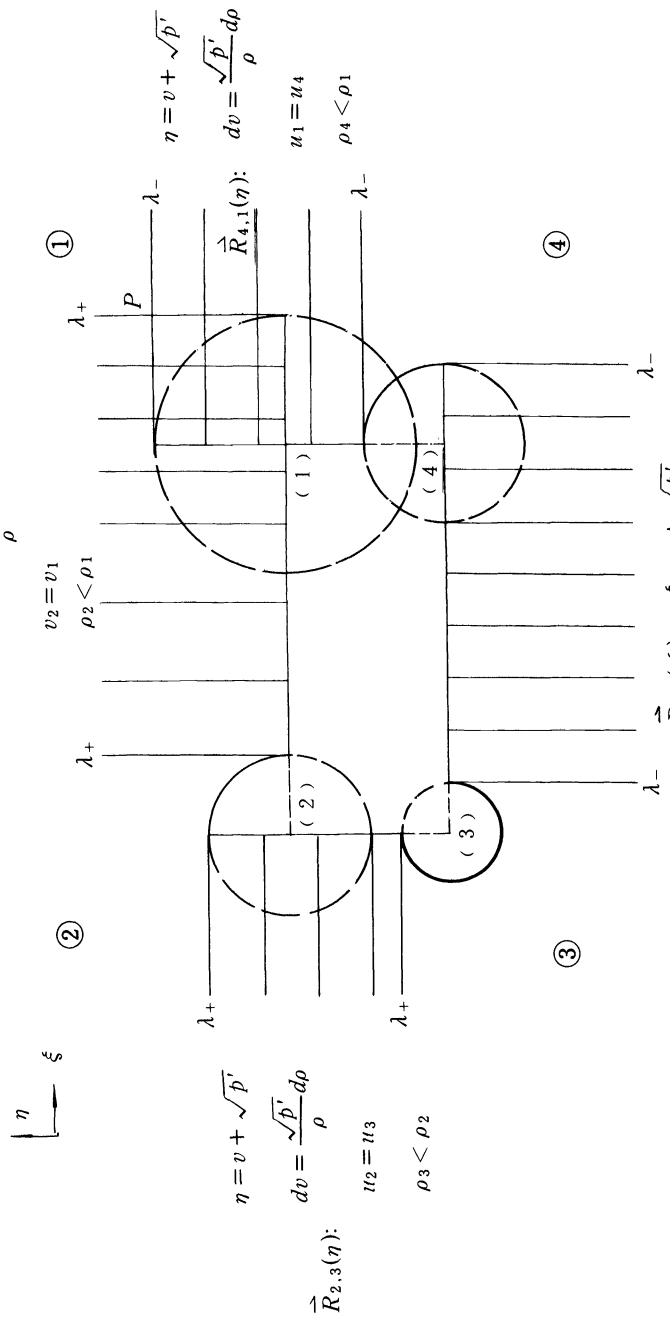

η
ξ
①
②
③
④
P
λ+
λ−
(1)
(2)
(3)
(4)
$\overleftarrow{R}_{1,2}(\xi)$: $\xi = u + \sqrt{p'}$, $du = \frac{\sqrt{p'}}{\rho}d\rho$
$v_2 = v_1$
$\rho_2 < \rho_1$
$\overleftarrow{R}_{4,1}(\eta)$: $\eta = v + \sqrt{p'}$, $dv = \frac{\sqrt{p'}}{\rho}d\rho$
$u_1 = u_4$
$\rho_4 < \rho_1$
$\overrightarrow{R}_{2,3}(\eta)$: $\eta = v + \sqrt{p'}$, $dv = \frac{\sqrt{p'}}{\rho}d\rho$
$w_2 = w_3$
$\rho_3 < \rho_2$
$\overrightarrow{R}_{3,4}(\xi)$: $\xi = u + \sqrt{p'}$, $du = \frac{\sqrt{p'}}{\rho}d\rho$
$v_3 = v_4$
$\rho_3 < \rho_4$

FIG. 4.1

we have

$$\int_{\rho_3}^{\rho_4} \frac{\sqrt{p'}}{\rho}\, d\rho = \int_{\rho_2}^{\rho_1} \frac{\sqrt{p'}}{\rho}\, d\rho, \qquad \int_{\rho_3}^{\rho_2} \frac{\sqrt{p'}}{\rho}\, d\rho = \int_{\rho_4}^{\rho_1} \frac{\sqrt{p'}}{\rho}\, d\rho.$$

Obviously, they are equivalent, and we call them compatibility conditions.

The system is strictly hyperbolic at infinity. We consider infinity as a Cauchy support. The data at the support are four constant states with four jumps. Solving them, we obtain a continuous rarefactive solution consisting of four constant states and four $\vec{R}$'s in the neighborhood of infinity. Extending the solution along stream lines, we should stop at either point P where two $\vec{R}$'s meet as $\vec{R}_{1,2}(\xi)$, ① and $\vec{R}_{4,1}(\eta)$, or the sonic curves as $\vec{R}_{2,3}(\eta)$, ③ and $\vec{R}_{4,3}(\xi)$ (Fig. 4.1). $\vec{R}_{1,2}(\xi)$ and $\vec{R}_{4,1}(\eta)$ meet at P before they reach their sonic curves. The boundary of $\vec{R}_{1,2}(\xi)$ should be the λ_- characteristic line extending from P. It penetrates $\vec{R}_{1,2}(\xi)$ and goes into ② with positive slope, and therefore it goes into $\vec{R}_{2,3}(\eta)$ without intersecting its sonic stem and then goes into ③ (Fig. 4.2). Analogously, we can extend the λ_+ characteristic line from P to go through $\vec{R}_{4,1}(\eta)$, ④ and $\vec{R}_{3,4}(\xi)$, and into ③.

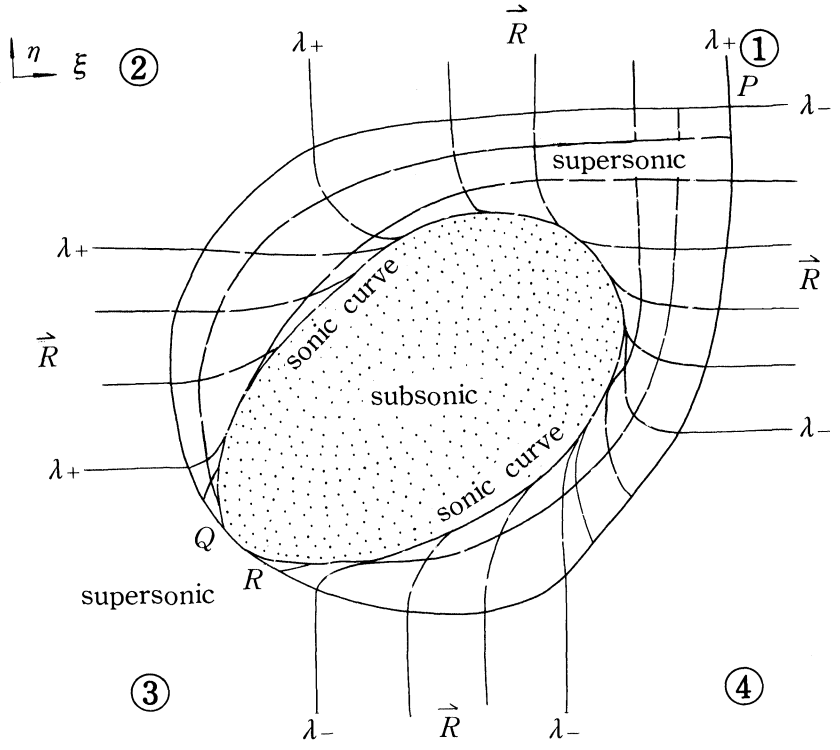


FIG. 4.2

There are two cases:

(a) These two characteristic lines extending from P become tangent to the sonic circle of ③ at Q and R, respectively, before they interact (Fig. 4.2). The λ_- characteristic line $\widehat{PQ}$ and the λ_+ characteristic line $\widehat{PR}$ form a Goursat problem. Q and R, the ends of the supports, are sonic points. We conjecture that this Goursat problem has a unique supersonic rarefactive continuous solution up to a sonic curve connecting Q and R

(but different from the arc $\widehat{QR}$) as its boundary. This sonic boundary, which is an envelope of the $\lambda_{\pm}$ characteristic lines, and the arc $\widehat{QR}$ of the sonic circle of ③ bound a domain. In this domain we conjecture that there exists a unique global subsonic continuous solution with a node of stream lines in it.

(b) Two λ characteristic lines intersect before they contact the sonic circle of ③. Our conjecture concerning the configuration of the solution is shown in Fig. 4.3.

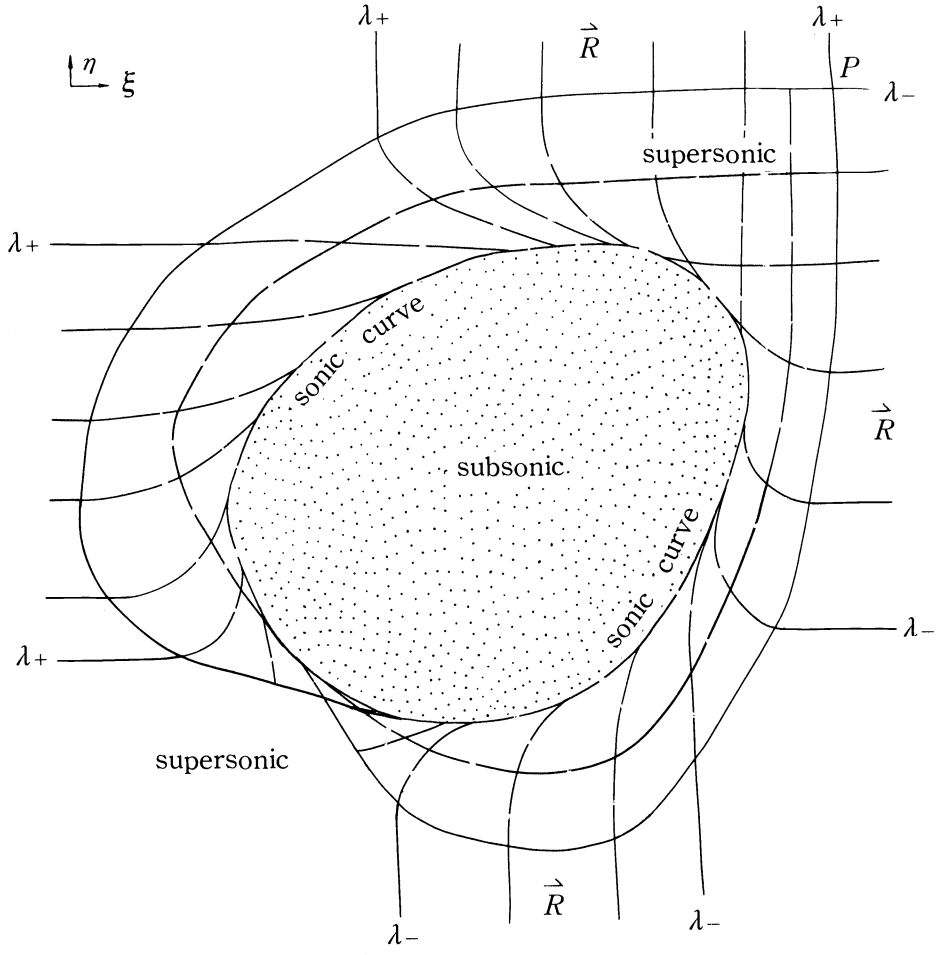


FIG. 4.3

There is an internal boundary, outside which the flow is supersonic and rarefactive. The boundary is the envelope of the λ characteristic lines. Inside the boundary, the flow is subsonic and rarefactive with a node of stream lines.

(2) Two $\vec{R}$'s and two $\tilde{R}$'s. It is pictured in Fig. 4.4. The inequalities of ρ_i are

$$\rho_3 > \frac{\rho_2}{\rho_4} < \rho_1.$$

The compatibility conditions are

$$\int_{\rho_2}^{\rho_3} \frac{\sqrt{p'}}{\rho}\, d\rho = \int_{\rho_4}^{\rho_1} \frac{\sqrt{p'}}{\rho}\, d\rho, \qquad \int_{\rho_2}^{\rho_1} \frac{\sqrt{p'}}{\rho}\, d\rho = \int_{\rho_4}^{\rho_3} \frac{\sqrt{p'}}{\rho}\, d\rho.$$

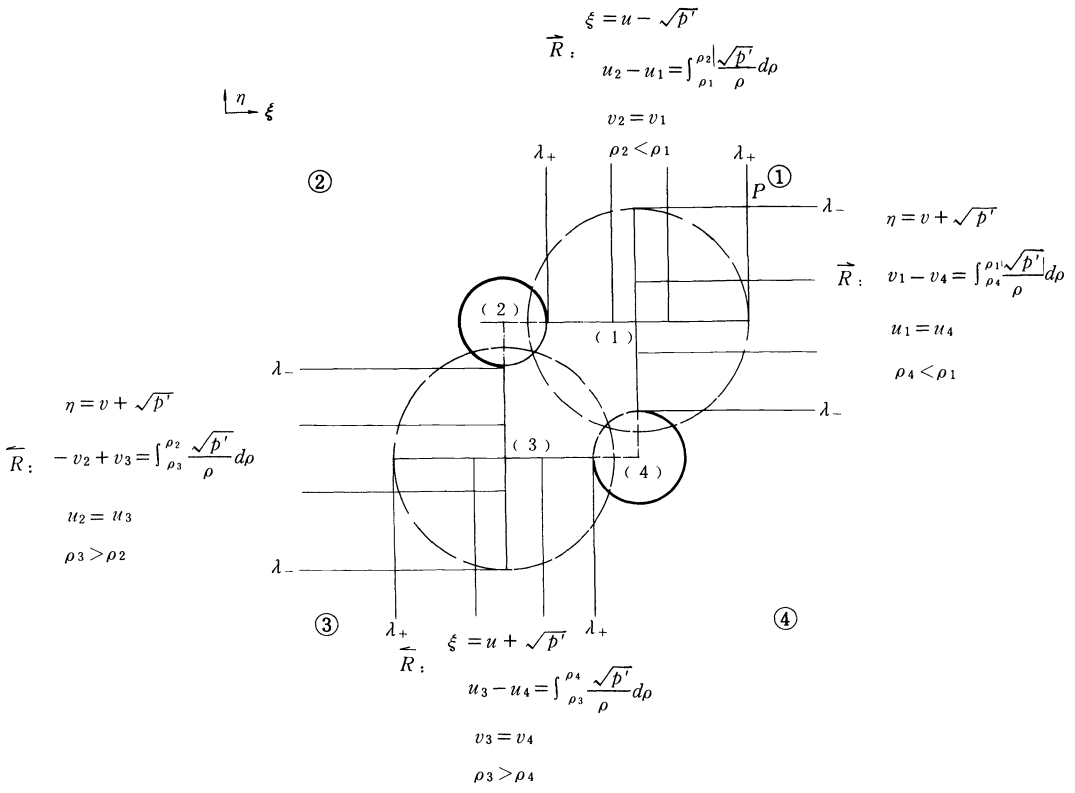


FIG. 4.4

It is easy to know from them that $\rho_2 = \rho_4$, $\rho_1 = \rho_3$; then $u_1 - u_2 = v_1 - v_4$. Thus Fig. 4.4 is symmetric to $\eta - \xi = -u_1 + v_1$ and $\xi + \eta = u_2 + v_2$. $\vec{R}_{1,2}$ and $\vec{R}_{1,4}$ meet at P. $\tilde{R}_{2,3}$ and $\tilde{R}_{4,3}$ meet at P'. Extend the following:

λ_- characteristic line from P to go through $\vec{R}_{1,2}$ and into ②,
λ_+ characteristic line from P to go through $\vec{R}_{1,4}$ and into ④,
λ_- characteristic line from P' to go through $\tilde{R}_{2,3}$ and into ②,
λ_+ characteristic line from P' to go through $\tilde{R}_{4,3}$ and into ④.

The $\lambda_\pm$ characteristic lines from P and P' in ② are either tangent to the sonic circle of ② before they meet, or they meet before they are tangent to the sonic circle of ②. The situation is analogous to ④. The configuration of solutions are conjectured as shown in Figs. 4.5 and 4.6.

There is a simple correspondence between two-dimensional pseudostationary flow and one-dimensional unstationary flow as follows. The stream line corresponds to $x =$ constant in the one-dimensional case where x is a Lagrangian coordinate. Infinity corresponds to $t = 0$ in the one-dimensional case, and the node of stream lines corresponds to $t = \infty$ in the one-dimensional case.

In addition, we have two conjectures for general pseudostationary flow as follows: (1) The pseudostationary flow is continuous on the whole plane if and only if it is continuous and rarefactive in a neighborhood of infinity. (2) The pseudostationary flow is smooth on the whole plane if and only if it is a constant state.

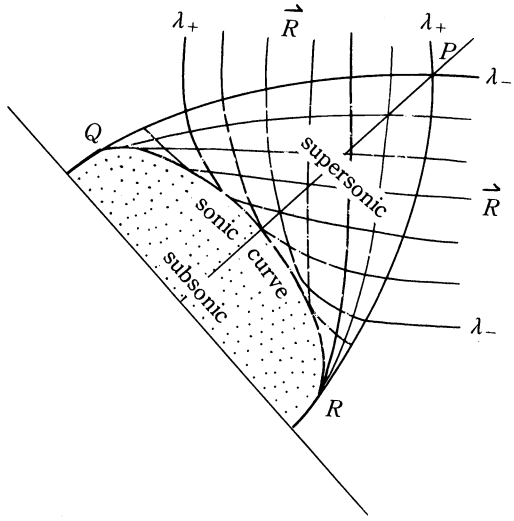


FIG. 4.5

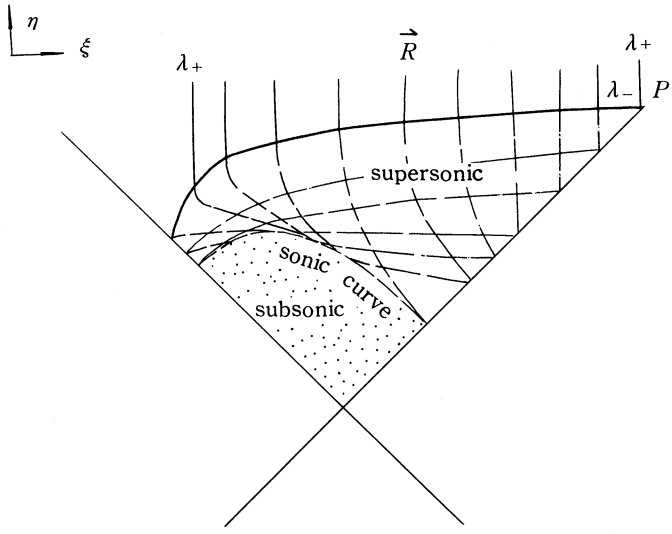


FIG. 4.6

5. Four shocks. There are two cases.

(1) Four $\vec{S}$'s (Fig. 5.1). The entropy conditions are

$$\rho_3 > \frac{\rho_2}{\rho_4} > \rho_1.$$

The compatibility conditions are

$$\sqrt{\frac{1}{\rho_1\rho_2}\,p'_{12}(\rho_2-\rho_1)} = \sqrt{\frac{1}{\rho_3\rho_4}\,p'_{34}(\rho_3-\rho_4)},$$

$$\sqrt{\frac{1}{\rho_4\rho_1}\,p'_{41}(\rho_4-\rho_1)} = \sqrt{\frac{1}{\rho_3\rho_2}\,p'_{32}(\rho_3-\rho_2)}.$$

They are not equivalent.

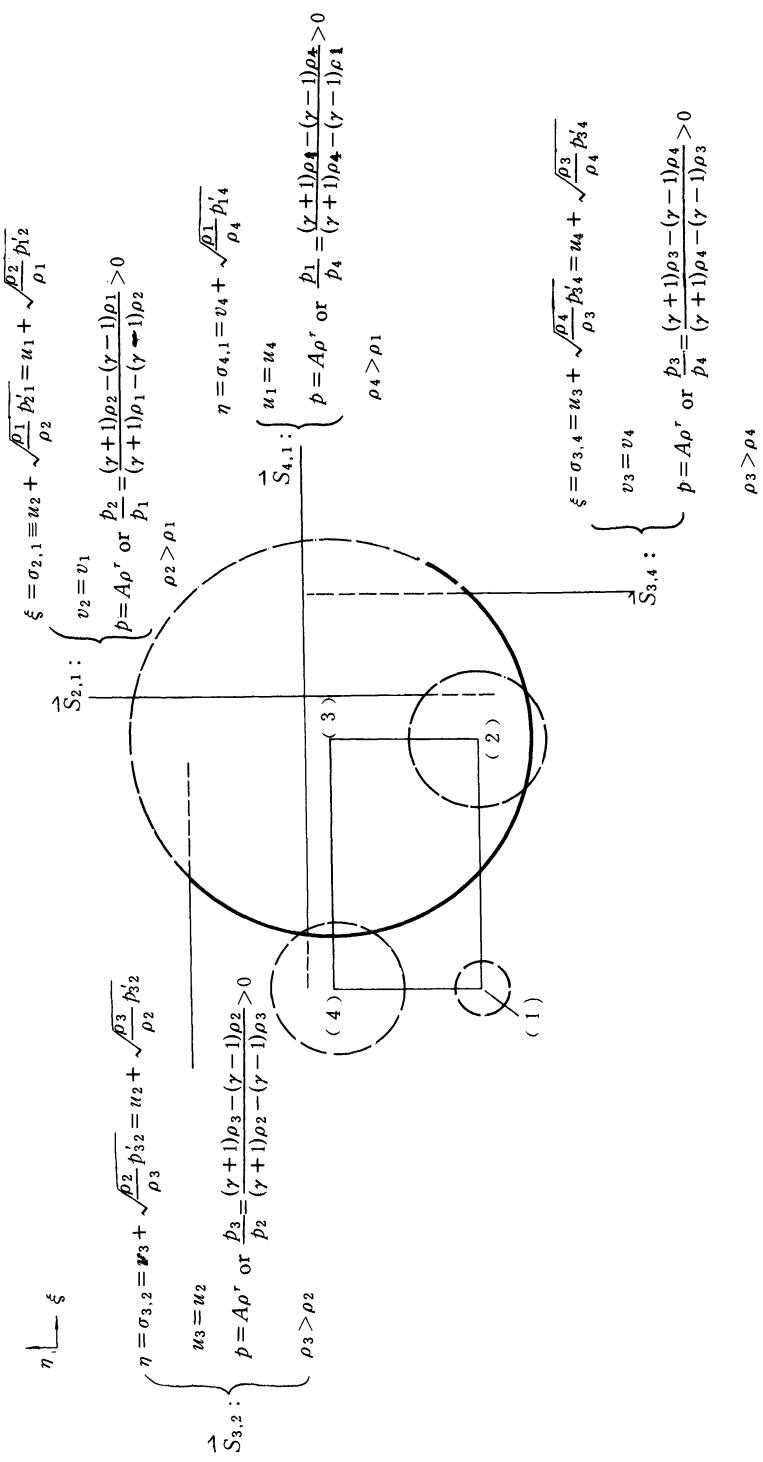

η
ξ
$\hat{S}_{2,1}$:
$\xi = \sigma_{2,1} \equiv u_2 + \sqrt{\frac{\rho_1}{\rho_2}}\, p'_{21} = u_1 + \sqrt{\frac{\rho_2}{\rho_1}}\, p'_{12}$
$v_2 = v_1$
$p = A\rho^r$ or $\frac{p_2}{p_1} = \frac{(\gamma+1)\rho_2 - (\gamma-1)\rho_1}{(\gamma+1)\rho_1 - (\gamma-1)\rho_2} > 0$
$\rho_2 > \rho_1$
$\hat{S}_{4,1}$:
$\eta = \sigma_{4,1} = v_4 + \sqrt{\frac{\rho_1}{\rho_4}}\, p'_{14}$
$u_1 = u_4$
$p = A\rho^r$ or $\frac{p_1}{p_4} = \frac{(\gamma+1)\rho_1 - (\gamma-1)\rho_4}{(\gamma+1)\rho_4 - (\gamma-1)\rho_1} > 0$
$\rho_4 > \rho_1$
$\hat{S}_{3,2}$:
$\eta = \sigma_{3,2} = v_3 + \sqrt{\frac{\rho_2}{\rho_3}}\, p'_{32} = u_2 + \sqrt{\frac{\rho_3}{\rho_2}}\, p'_{32}$
$u_3 = u_2$
$p = A\rho^r$ or $\frac{p_3}{p_2} = \frac{(\gamma+1)\rho_3 - (\gamma-1)\rho_2}{(\gamma+1)\rho_2 - (\gamma-1)\rho_3} > 0$
$\rho_3 > \rho_2$
$\hat{S}_{3,4}$:
$\xi = \sigma_{3,4} = u_3 + \sqrt{\frac{\rho_4}{\rho_3}}\, p'_{34} = u_4 + \sqrt{\frac{\rho_3}{\rho_4}}\, p'_{34}$
$v_3 = v_4$
$p = A\rho^r$ or $\frac{p_3}{p_4} = \frac{(\gamma+1)\rho_3 - (\gamma-1)\rho_4}{(\gamma+1)\rho_4 - (\gamma-1)\rho_3} > 0$
$\rho_3 > \rho_4$
(1)
(2)
(3)
(4)

FIG. 5.1

We can prove that

$$\sigma_{1,2}<\sigma_{3,4}, \qquad \sigma_{1,4}<\sigma_{2,3}.$$

In fact, it is equivalent to proving that

$$\frac{\rho_1}{\rho_2}p'_{12}<\frac{\rho_4}{\rho_3}p'_{34}, \qquad \frac{\rho_1}{\rho_4}p'_{14}<\frac{\rho_2}{\rho_3}p'_{23}.$$

Making use of the compatibility conditions, we see this is equivalent to proving that

$$\rho_1\rho_3<\rho_2\rho_4.$$

Put

$$\rho_3=x\rho_1, \quad \rho_2=y\rho_1, \quad \rho_4=z\rho_1;$$

then our problem reduces to proving that

$$x<yz$$

under entropy conditions

$$x>\frac{y}{z}>1$$

and compatibility conditions

$$\frac{(x^\gamma-z^\gamma)(x-z)}{xz}=\frac{(y^\gamma-1)(y-1)}{y} \quad \text{(isentropic)}$$

or

$$\frac{(x-z)^2}{xz}\cdot\frac{(\gamma+1)z-(\gamma-1)}{(\gamma+1)z-(\gamma-1)x}=\frac{(y-1)^2}{y}\cdot\frac{(\gamma+1)-(\gamma-1)z}{(\gamma+1)-(\gamma-1)y} \quad \text{(adiabatic)}$$

with

$$(\gamma+1)-(\gamma-1)y>0, \qquad (\gamma+1)-(\gamma-1)z>0.$$

For isentropic flow, put

$$f(x)\equiv(x^\gamma-z^\gamma)(x-z)y-xz(y^\gamma-1)(y-1).$$

It is easy to check that

$$F(z)<0,$$
$$f(yz)=y(y^\gamma-1)(y-1)z^2(z^{\gamma-1}-1)>0,$$
$$f''(x)=\gamma yx^{\gamma-2}[(\gamma+1)x-(\gamma-1)z]>0.$$

So, when $f(x)=0$, we have $x<yz$.

For adiabatic flow, put

$$f(x)\equiv(x-z)^2[(\gamma+1)z-(\gamma-1)]y[(\gamma+1)-(\gamma-1)y]$$
$$-xz[(\gamma+1)z-(\gamma-1)x](y-1)^2[(\gamma+1)-(\gamma-1)z].$$

It is easy to check that

$$f(z)<0,$$
$$f(yz)=yz^2(y-1)^2[(\gamma+1)-(\gamma-1)y](\gamma-1)(z^2-1)>0,$$
$$f''(x)=(\gamma-1)z(y-1)^2[(\gamma+1)-(\gamma-1)z]$$
$$+2[(\gamma+1)z-(\gamma-1)]y[(\gamma+1)-(\gamma-1)y]>0.$$

So, when $f(x)=0$, we have $x<yz$. □

Thus, $\vec{S}_{2,3}(\eta)$ intersects the sonic circle of ③ before it meets other shocks. The same is true for $\vec{S}_{3,4}(\xi)$. $\vec{S}_{2,3}(\eta)$, ③ and $\vec{S}_{3,4}(\xi)$ should stop at the sonic circle of ③ (Fig. 5.2). Across the circle, the flow should be subsonic and the shocks should be

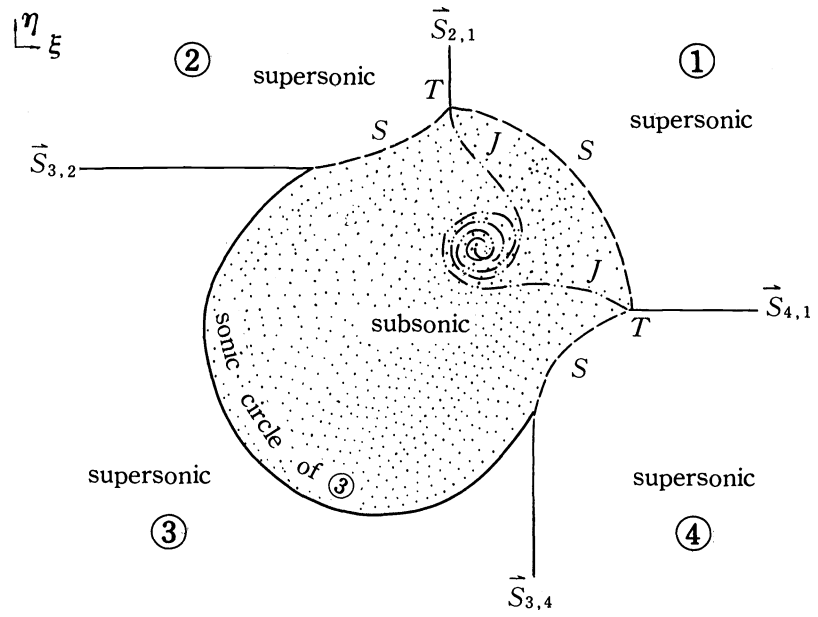


FIG. 5.2

bent to intersect the other two shocks $\vec{S}_{2,1}(\xi)$ and $\vec{S}_{1,4}(\eta)$, respectively. The curved shock with ② on one side must be convex with respect to (u_2, v_2). The intersection points T of shocks should have a triple shock configuration. The shocks from two triple points T will match together smoothly. The flow behind them should be subsonic and, in front of them, supersonic. The slip lines coming from the T's are linear jumps. There are no entropy conditions for slip lines. They should end with spirals in the subsonic flow. The subsonic flow and its free boundary of shocks as well as the triple points should be determined simultaneously.

(2) Two $\vec{S}$'s and two $\tilde{S}$'s (Fig. 5.3). The entropy conditions are

$$\rho_1 \lessgtr \frac{\rho_2}{\rho_4} \gtrless \rho_3.$$

The compatibility conditions are

$$\sqrt{\frac{1}{\rho_1\rho_2}p'_{12}}(\rho_1-\rho_2)=\sqrt{\frac{1}{\rho_3\rho_4}p'_{34}}(\rho_3-\rho_4), \qquad \sqrt{\frac{1}{\rho_2\rho_3}p'_{23}}(\rho_3-\rho_2)=\sqrt{\frac{1}{\rho_1\rho_4}p'_{14}}(\rho_1-\rho_4)$$

from which we know that $\rho_1=\rho_3$, $\rho_2=\rho_4$; then $u_2-u_1=v_3-v_2$ and Fig. 5.3 is symmetric to $\xi+\eta=u_2+v_2$ and $\xi-\eta=u_1-v_1$.

Obviously, $\sigma_{1,2}>\sigma_{3,4}$ and $\sigma_{1,4}>\sigma_{2,3}$, so we consider the intersection point P of $\vec{S}_{1,2}$ and $\vec{S}_{1,4}$. Both ② and ④ may be subsonic, sonic, or supersonic with respect to P. If ② and ④ are subsonic at P, we should go back to their sonic circle, and $\vec{S}_{1,2}$ and $\vec{S}_{1,4}$ should be bent from there. We conjecture the structure of the solution as in Fig. 5.4.

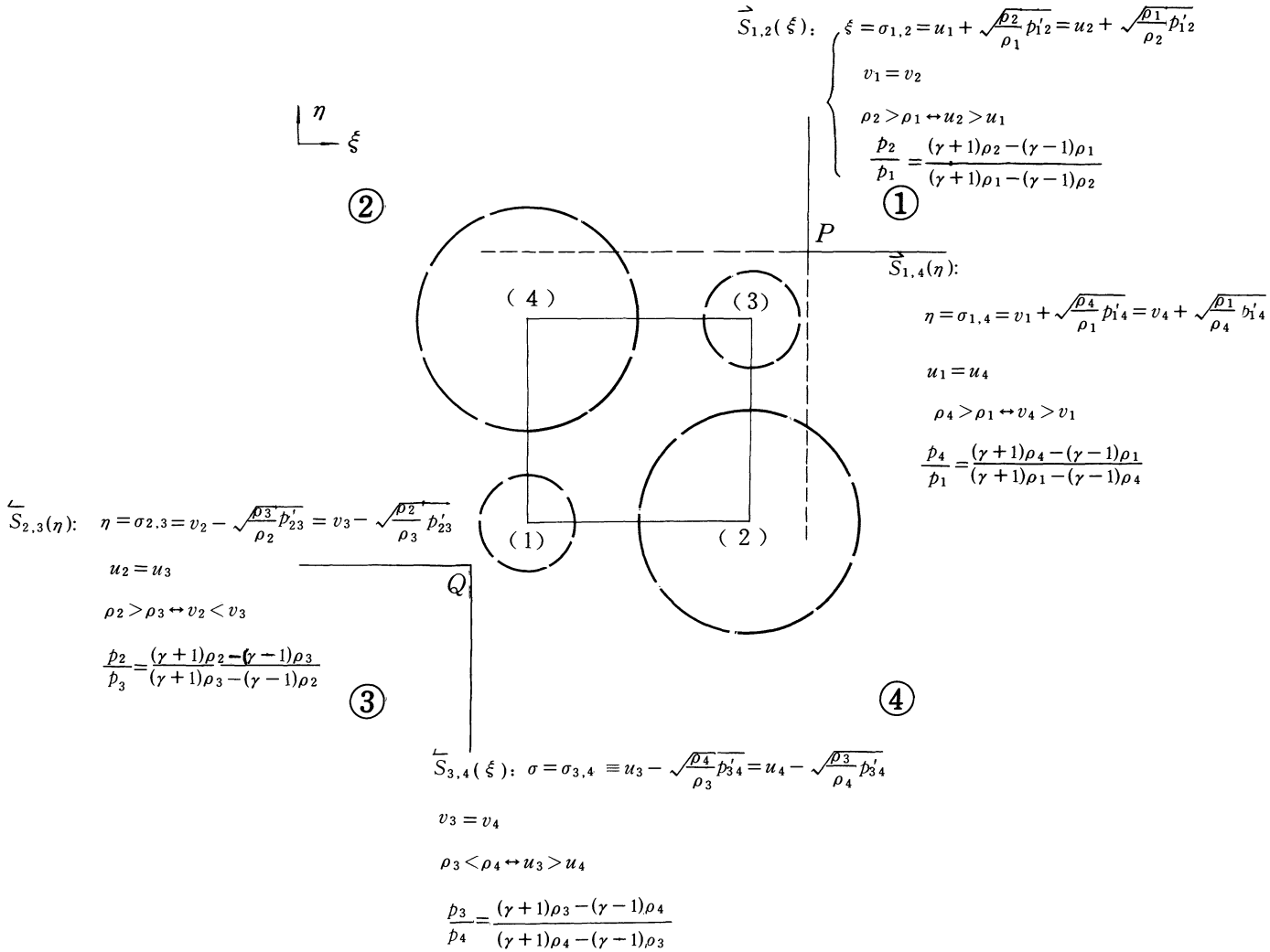

$\overrightarrow{S}_{1,2}(\xi)$: $\xi=\sigma_{1,2}=u_1+\sqrt{\frac{p_2}{\rho_1}}p'_{12}=u_2+\sqrt{\frac{p_1}{\rho_2}}p'_{12}$
$v_1=v_2$
$\rho_2>\rho_1\leftrightarrow u_2>u_1$
$\frac{p_2}{p_1}=\frac{(\gamma+1)\rho_2-(\gamma-1)\rho_1}{(\gamma+1)\rho_1-(\gamma-1)\rho_2}$
η
ξ
②
①
P
$\overrightarrow{S}_{1,4}(\eta)$:
$\eta=\sigma_{1,4}=v_1+\sqrt{\frac{p_4}{\rho_1}}p'_{14}=v_4+\sqrt{\frac{p_1}{\rho_4}}p'_{14}$
$u_1=u_4$
$\rho_4>\rho_1\leftrightarrow v_4>v_1$
$\frac{p_4}{p_1}=\frac{(\gamma+1)\rho_4-(\gamma-1)\rho_1}{(\gamma+1)\rho_1-(\gamma-1)\rho_4}$
(4)
(3)
(1)
(2)
$\overleftarrow{S}_{2,3}(\eta)$: $\eta=\sigma_{2,3}=v_2-\sqrt{\frac{p_3}{\rho_2}}p'_{23}=v_3-\sqrt{\frac{p_2}{\rho_3}}p'_{23}$
$u_2=u_3$
$\rho_2>\rho_3\leftrightarrow v_2<v_3$
$\frac{p_2}{p_3}=\frac{(\gamma+1)\rho_2-(\gamma-1)\rho_3}{(\gamma+1)\rho_3-(\gamma-1)\rho_2}$
Q
③
④
$\overleftarrow{S}_{3,4}(\xi)$: $\sigma=\sigma_{3,4}\equiv u_3-\sqrt{\frac{p_4}{\rho_3}}p'_{34}=u_4-\sqrt{\frac{p_3}{\rho_4}}p'_{34}$
$v_3=v_4$
$\rho_3<\rho_4\leftrightarrow u_3>u_4$
$\frac{p_3}{p_4}=\frac{(\gamma+1)\rho_3-(\gamma-1)\rho_4}{(\gamma+1)\rho_4-(\gamma-1)\rho_3}$

FIG. 5.3

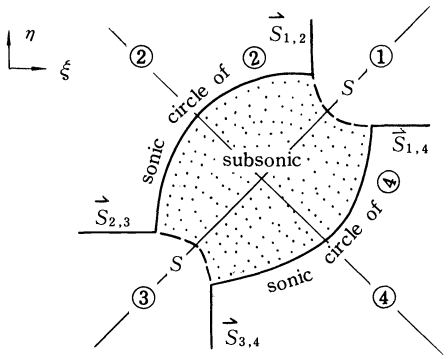


FIG. 5.4

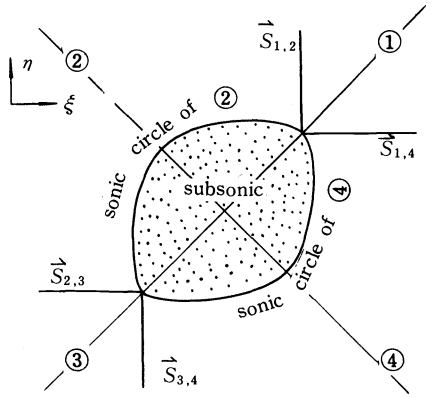


FIG. 5.5

If ② and ④ are sonic at P, the structure of the solution should be as shown in Fig. 5.5.

If ② and ④ are supersonic at P, the problem is just the same as the reflection of an oblique shock in steady flow at P, since the straight line through P and Q is a stream line. But it is not the diffraction of a planar shock around a compressive corner because the two symmetric axes, two stream lines, are perpendicular to each other. By the way, it is easy to prove that ② and ④ are supersonic at P when $1<\gamma\leqq 3$.

There are two possibilities as follows.

In the case of adiabatic flow, if

$$(*)\qquad \sqrt[3]{-\frac{q}{2}+\sqrt{\frac{q^2}{4}+\frac{p^3}{27}}}+\sqrt[3]{-\frac{q}{2}+\sqrt{\frac{q^2}{4}+\frac{p^3}{27}}}\leqq 2,$$

where

$$p=\frac{3ac-b^2}{3a^2}<0,\quad q=\frac{2b^3-9abc+27a^2d}{27a^3}<0,\quad a=\zeta^2>0,$$

$$b=-\zeta\{(3\zeta-2)+(\gamma-1)(\zeta-1)[(\gamma+1)(\zeta-1)+2]\}<0,$$

$$c=-(\zeta-1)[2(\gamma+1)\zeta(\zeta-1)+(\zeta+1)]<0,$$

$$d=-(\zeta-1)^2<0,\qquad \zeta=\rho_2/\rho_1,$$

the structure of the solution is regular reflection [12] (Fig. 5.6).

If the state (denoted by ⑤) behind the reflection shock at P is subsonic or sonic, there is a curved reflection shock that connects the supersonic flow and subsonic flow, otherwise the reflection shock is a straight line until it reaches the sonic circle of constant state ⑤. Then the subsonic flow should be bounded by the sonic circle of ⑤, the curved reflection shock S, as well as the two perpendicular stream lines.

If condition (*) is violated, the structure of the solution should be Mach reflections as shown in Fig. 5.7. There are several subcases: simple Mach reflection if ⑤ is sonic or subsonic; complex Mach reflection if ⑤ is supersonic and the sonic circle of ⑤ intersects with J; and double Mach reflection if ⑤ is supersonic but the sonic circle of ⑤ does not intersect with J. It seems that there is no spiral because of symmetry, but the spiral should appear after perturbation [13]-[15] (Fig. 5.8).

This phenomenon reflects the instability of two-dimensional flow.

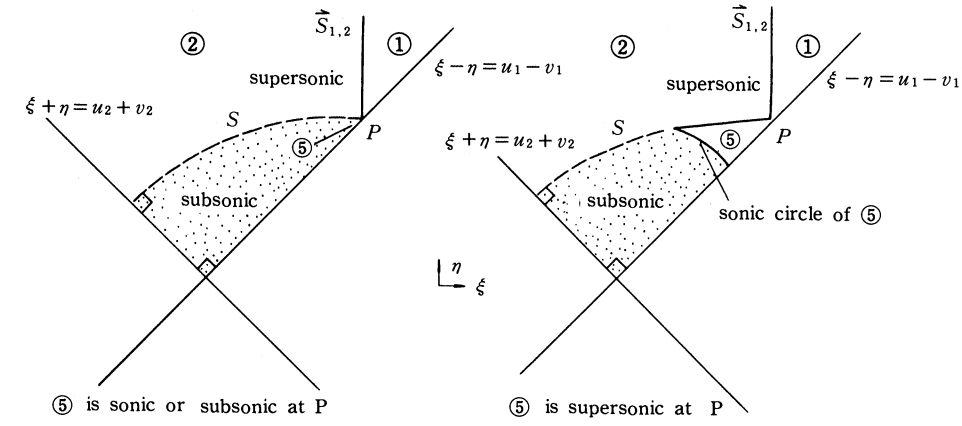


FIG. 5.6

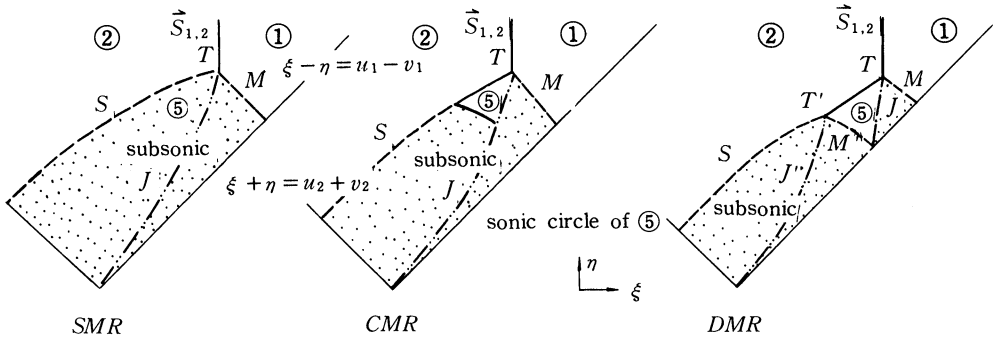


FIG. 5.7

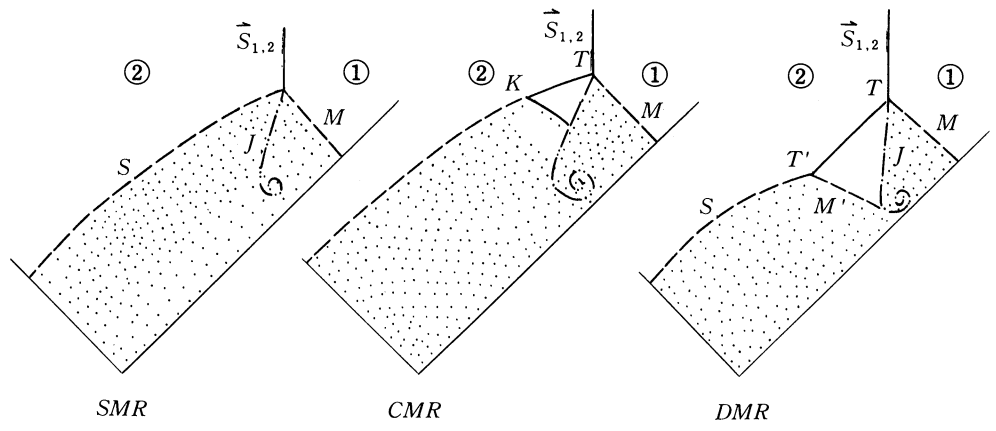


FIG. 5.8

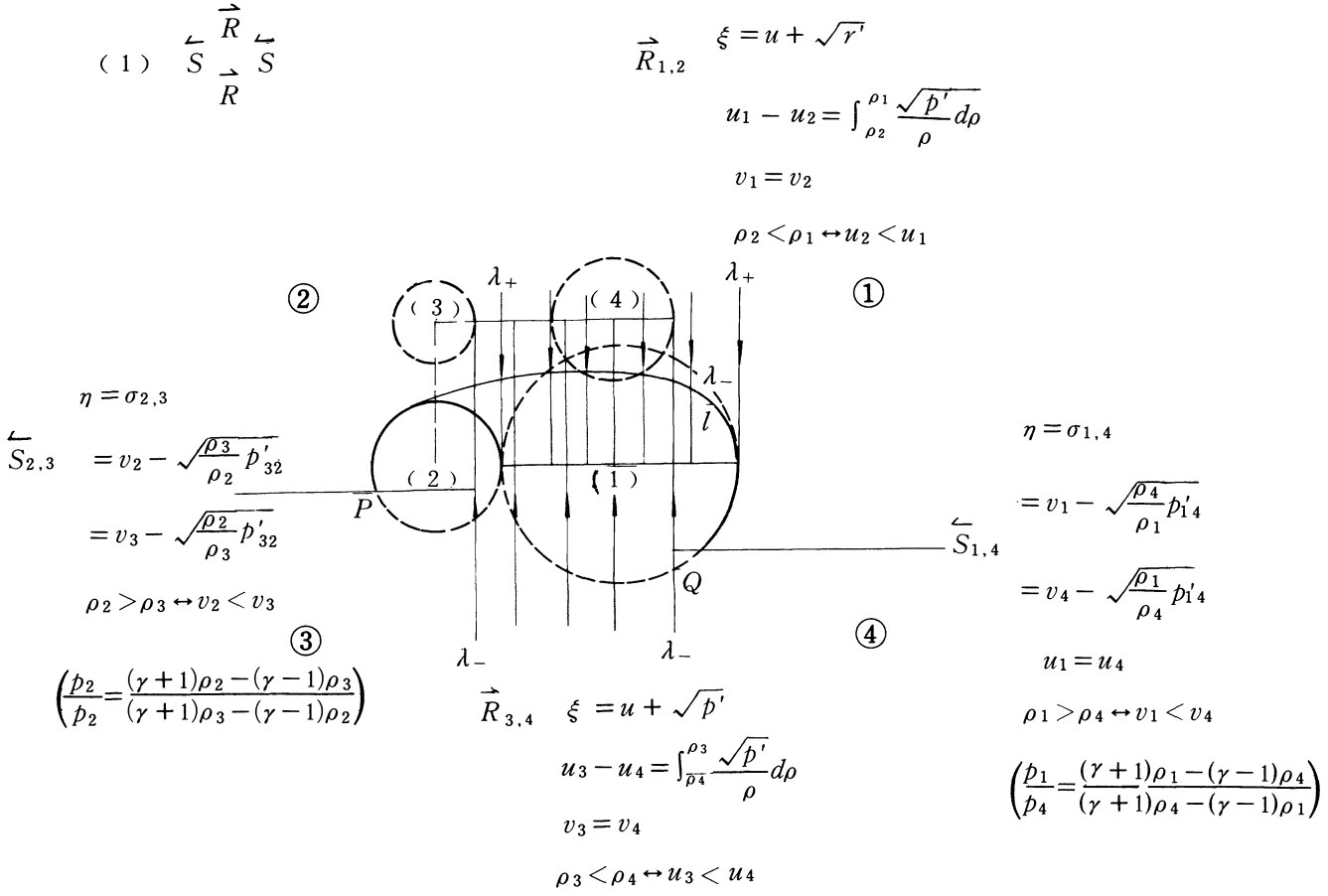

(1) $\overleftarrow{S}$ $\vec{R}$ $\overleftarrow{S}$ $\vec{R}$
$\vec{R}_{1,2}$ $\xi = u + \sqrt{r'}$
$u_1 - u_2 = \int_{\rho_2}^{\rho_1} \frac{\sqrt{p'}}{\rho} d\rho$
$v_1 = v_2$
$\rho_2 < \rho_1 \leftrightarrow u_2 < u_1$
②
①
③
④
λ_+
λ_-
(1)
(2)
(3)
(4)
l
P
Q
$\eta = \sigma_{2,3}$
$\overleftarrow{S}_{2,3} = v_2 - \sqrt{\frac{\rho_3}{\rho_2} p'_{32}}$
$= v_3 - \sqrt{\frac{\rho_2}{\rho_3} p'_{32}}$
$\rho_2 > \rho_3 \leftrightarrow v_2 < v_3$
$\left(\frac{p_2}{p_2} = \frac{(\gamma+1)\rho_2 - (\gamma-1)\rho_3}{(\gamma+1)\rho_3 - (\gamma-1)\rho_2}\right)$
$\eta = \sigma_{1,4}$
$= v_1 - \sqrt{\frac{\rho_4}{\rho_1} p'_{14}}$
$= v_4 - \sqrt{\frac{\rho_1}{\rho_4} p'_{14}}$
$\overleftarrow{S}_{1,4}$
$u_1 = u_4$
$\rho_1 > \rho_4 \leftrightarrow v_1 < v_4$
$\left(\frac{p_1}{p_4} = \frac{(\gamma+1)\rho_1 - (\gamma-1)\rho_4}{(\gamma+1)\rho_4 - (\gamma-1)\rho_1}\right)$
$\vec{R}_{3,4}$ $\xi = u + \sqrt{p'}$
$u_3 - u_4 = \int_{\rho_4}^{\rho_3} \frac{\sqrt{p'}}{\rho} d\rho$
$v_3 = v_4$
$\rho_3 < \rho_4 \leftrightarrow u_3 < u_4$

FIG. 6.1

6. Two rarefaction waves and two shocks.

(1) $\vec{R}\ \tilde{S}\ \tilde{S}$ (Fig. 6.1). We have

$$\rho_3 < \begin{matrix}\rho_2 \\ \rho_4\end{matrix} < \rho_1$$

and

$$\int_{\rho_2}^{\rho_1} \frac{\sqrt{p'}}{\rho}\,d\rho = \int_{\rho_3}^{\rho_4} \frac{\sqrt{p'}}{\rho}\,d\rho, \qquad \sqrt{\frac{1}{\rho_3\rho_2}\,p'_{32}(\rho_2-\rho_3)} = \sqrt{\frac{1}{\rho_1\rho_4}\,p'_{14}(\rho_1-\rho_4)}.$$

$\tilde{S}_{2,3}$ must meet the sonic circle of ② first; then it is bent to continue on to penetrate $\vec{R}_{34}$, go through $\vec{R}_{43}$ to intersect $\tilde{S}_{14}$ and form a triple point. The boundary of the supersonic flow of $\vec{R}_{12}$ coming from infinity is the λ_- characteristic line (denoted by l in Fig. 6.1) through the end of the sonic stem of $\vec{R}_{12}$. The shock from the triple point should go through $\vec{R}_{12}$ above line l, then into ② and be tangent to the sonic circle of ② somewhere (Fig. 6.2).

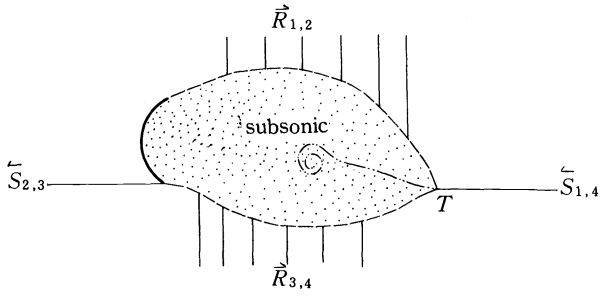


FIG. 6.2

(2) $\vec{R}\ \vec{S}\ \tilde{S}$ (Fig. 6.3). We have

$$\rho_3 > \begin{matrix}\rho_2 \\ \rho_4\end{matrix} < \rho_1$$

and

$$\int_{\rho_2}^{\rho_1} \frac{\sqrt{p'}}{\rho}\,d\rho = \int_{\rho_4}^{\rho_3} \frac{\sqrt{p'}}{\rho}\,d\rho, \qquad \sqrt{\frac{1}{\rho_3\rho_2}\,p'_{32}(\rho_3-\rho_2)} = \sqrt{\frac{1}{\rho_1\rho_4}\,p'_{14}(\rho_1-\rho_4)}.$$

The shock $\vec{S}_{2,3}$ must meet the sonic circle of ② first; then it penetrates $\vec{R}_{1,2}$ and goes into ① to intersect $\tilde{S}_{14}$ and forms a triple point. The new shock coming from the triple point continues on to penetrate $\vec{R}_{3,4}$ and enters ③ to intersect $\vec{S}_{2,3}$ and forms another triple point. This triple point on $\vec{S}_{2,3}$ may not be the sonic point we started with, so we modify the structure by starting at a (triple) point that is also the end point of the shock coming from $\tilde{S}_{1,4}$. Therefore we conjecture that the structure of the solution is as shown in Fig. 6.4.

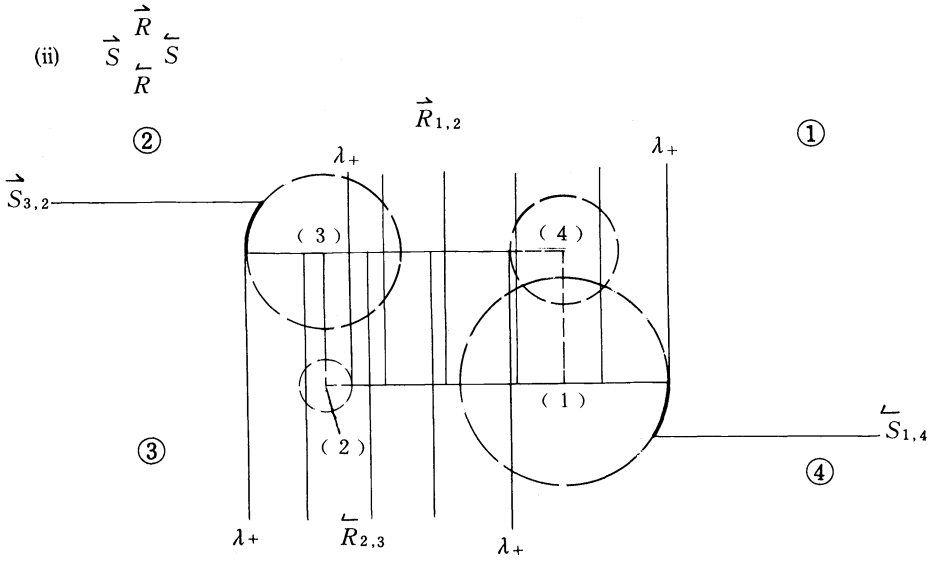


FIG. 6.3

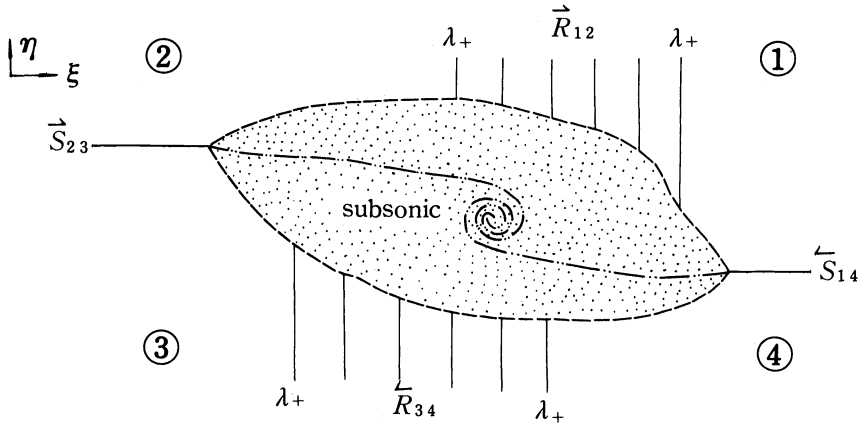


FIG. 6.4

7. The cases involving slip lines.

(1) Four J's. There are only two subcases (Fig. 7.1). For both we have

$$u_1 = u_2, \quad u_3 = u_4, \quad v_2 = v_3, \quad v_4 = v_1,$$

$$\rho_1 = \rho_2 = \rho_3 = \rho_4 \quad \text{(isentropic)} \quad \text{or} \quad p_1 = p_2 = p_3 = p_3 \quad \text{(adiabatic)}.$$

In the first subcase (Fig. 7.2), J_{41}, ④, and J_{34} should stop at the sonic circle of ④. It is easy to see that the slip line having constant state ③ at one side must be a straight line through point (3). So J_{34} cannot be bent to be a boundary between ③ and a subsonic flow. We conjecture that there is a free boundary of shock that starts tangentially to the λ-characteristic at the intersection point of $J_{3,4}$ and the sonic circle

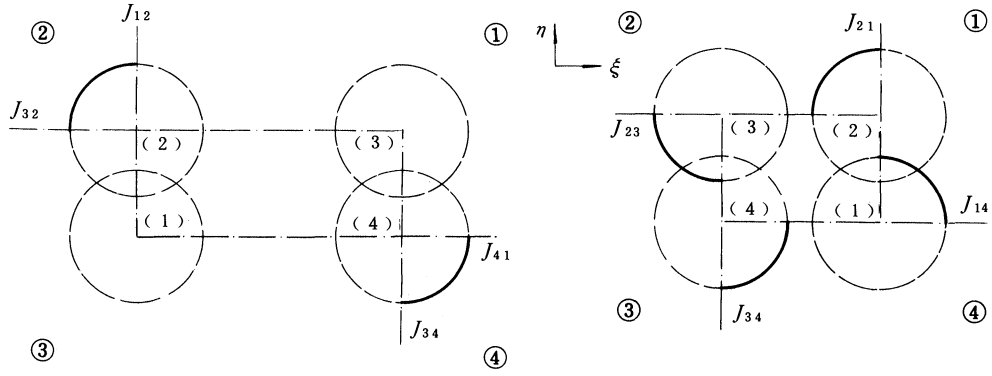


FIG. 7.1

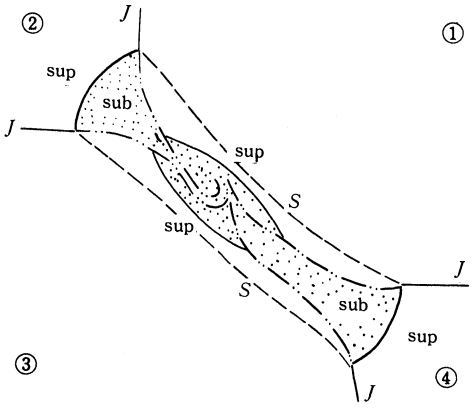


FIG. 7.2

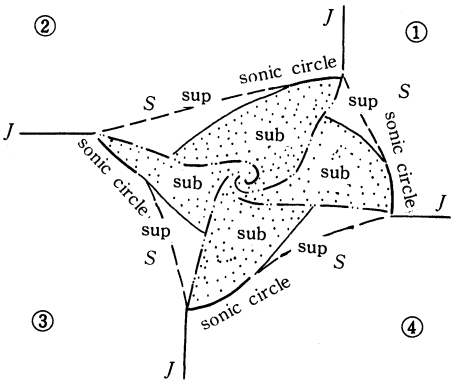


FIG. 7.3

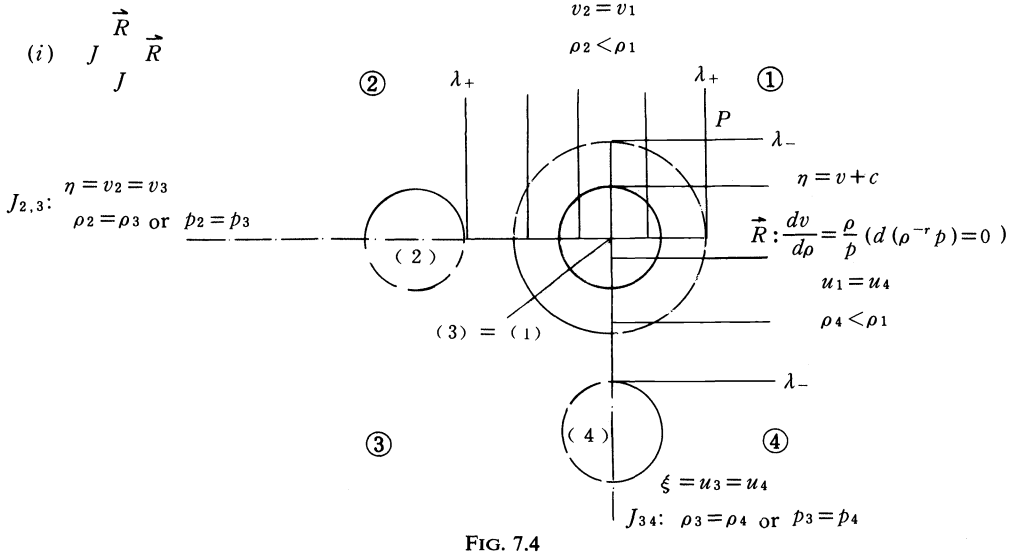


FIG. 7.4

of ④ and ends tangentially to the λ_+ characteristic at the intersection point of $J_{3,2}$ and the sonic circle of ②. This shock wave connects ③ to a nonconstant supersonic flow, which becomes subsonic in the central region gradually by itself or through the slip lines. The structure is similar in the upper right region. The slip lines end with spirals in the subsonic region.

In the second subcase, as the undetermined shocks should be out of their corresponding sonic circles and end tangentially at the circles, we conjecture that the solution is as shown in Fig. 7.3.

(2) Two J's.

(a) Two J's are neighbors.

(i) $J_J^{\vec{R}}\vec{R}$ (Fig. 7.4). Figure 7.4 is symmetric to $\xi-\eta=u_1-v_1$. It is a combination of Figs. 4.4 and 7.1. Our conjecture of the solution is shown in Fig. 7.5.

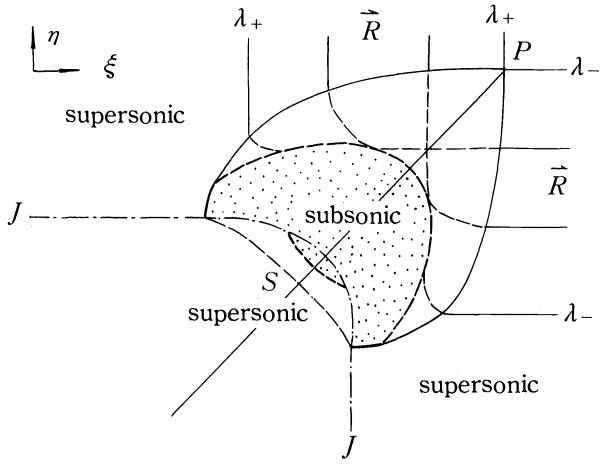


FIG. 7.5

(ii) $J_J^{\tilde{R}}\tilde{R}$ (Fig. 7.6). Figure 7.6 is symmetric to $\xi-\eta=u_1-v_1$. We have $\rho_1<\rho_2=\rho_3=\rho_4$, $(u_1, v_1)=(u_3, v_3)$. Solve $\tilde{R}_{12}$, ①, and $\tilde{R}_{14}$ to $\lambda_\mp$ through $p_\mp$, ① to $\widehat{Q_-Q_+}$. These curves and undetermined shock $\widehat{P_-P_+}$ bound a domain inside which there is a subsonic domain as shown in Fig. 7.7.

(iii) $J_J^{\vec{S}}\vec{S}$ (Fig. 7.8). We have $\rho_1=\rho_2=\rho_4>\rho_3$. Figure 7.8 is symmetric to $\xi-\eta=u_1-v_1$, and it is a combination of the upper right part of Fig. 5.3 and the lower left part of the first part of Fig. 7.1. The solutions are combinations of parts of Figs. 5.4–5.7 and 7.2.

(iv) $J_J^{\tilde{S}}\tilde{S}$ (Fig. 7.9). We have $\rho_1>\rho_2=\rho_3=\rho_4$. It is a combination of parts of Figs. 5.1 and 7.1. The solution is a combination of parts of Figs. 5.2 and 7.2.

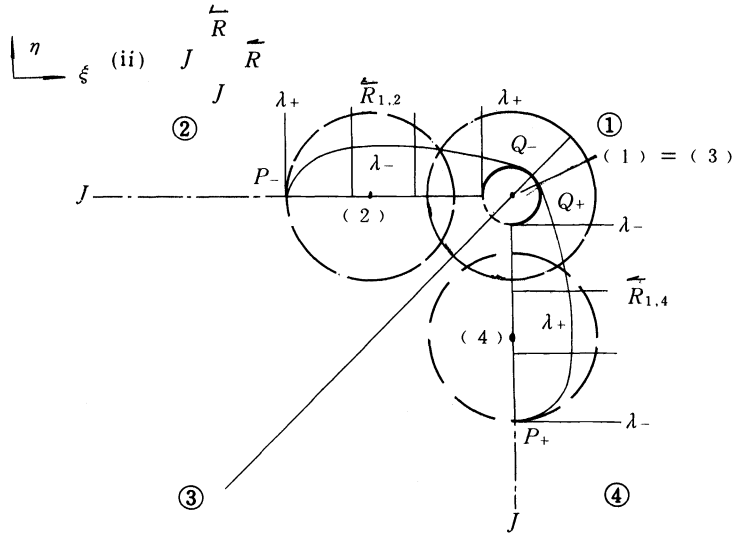


FIG. 7.6

(v) $J^{\vec{R}}_{J}\tilde{S}$ (Fig. 7.10). We have $\rho_2=\rho_3=\rho_4<\rho_1$ and $(u_1, v_1)=(u_3, v_3)$. It is a combination of parts of Figs. 6.1 and 7.1. The solution is a combination of parts of Figs. 6.2 and 7.2 (Fig. 7.11).

(vi) $J^{\vec{S}}_{J}\tilde{R}$ (Fig. 7.12). We have $\rho_1<\rho_2=\rho_3=\rho_4$. It is a combination of parts of Figs. 6.3 and 7.1. The solution (Fig. 7.13) is a combination of parts of Figs. 6.4 and 7.2.

(b) Two J's are not neighbors.

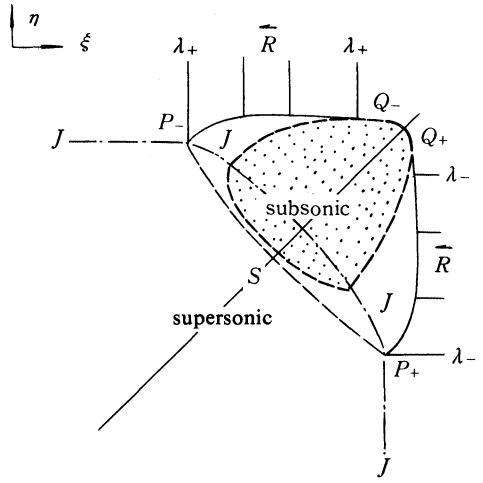


FIG. 7.7

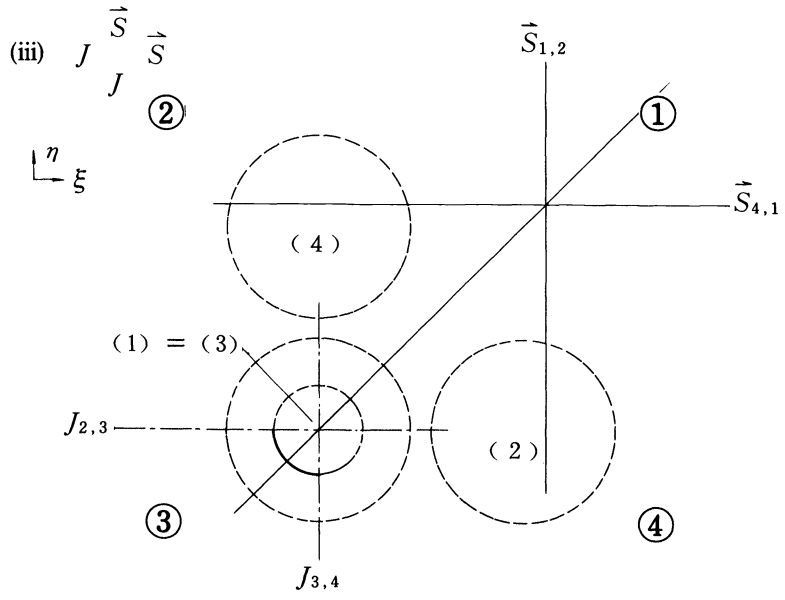


FIG. 7.8

(i) $\vec{R}\overset{J}{J}\vec{R}$ (Fig. 7.14). Cut the $\vec{R}$ along the sonic stem into two parts, then slip them up or down. ① should stop at point P. A shock should start from P continuously at the side of ② and be tangent to the λ_- characteristic there. The shock should be convex with respect to point (2), so it will go into $\vec{R}_{23}$ and then penetrate $\vec{R}_{23}$ into ③; then it will be tangent to the sonic circle of ③ and will vanish there continuously (Fig. 7.15). Analogously, ③ should stop at point Q. A shock should start from Q continuously

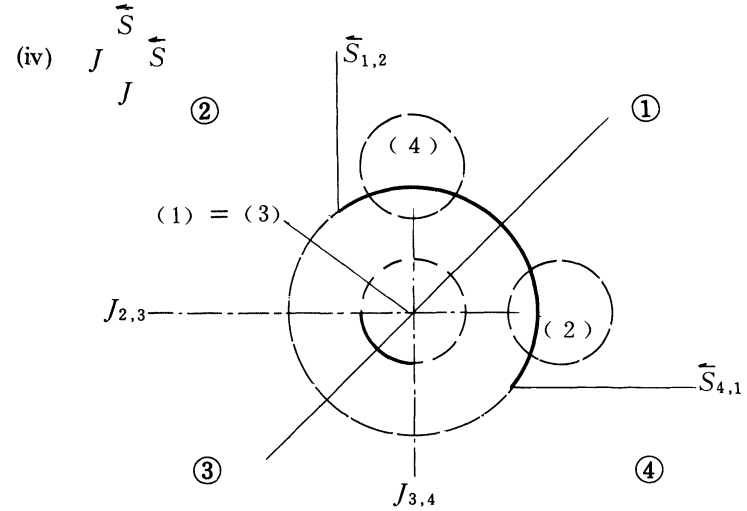


FIG. 7.9

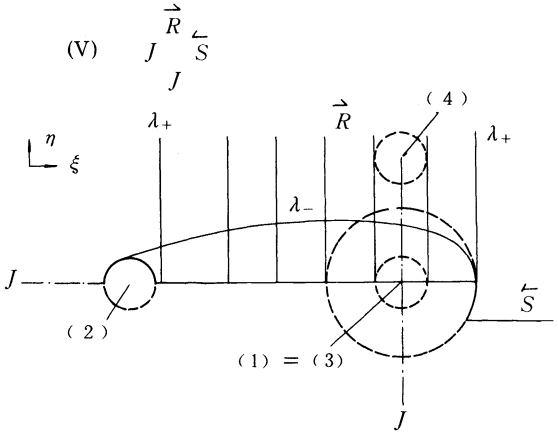


FIG. 7.10

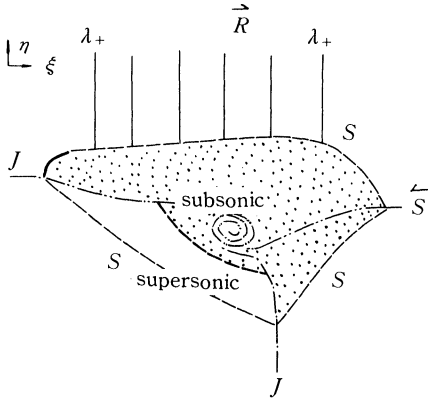


FIG. 7.11

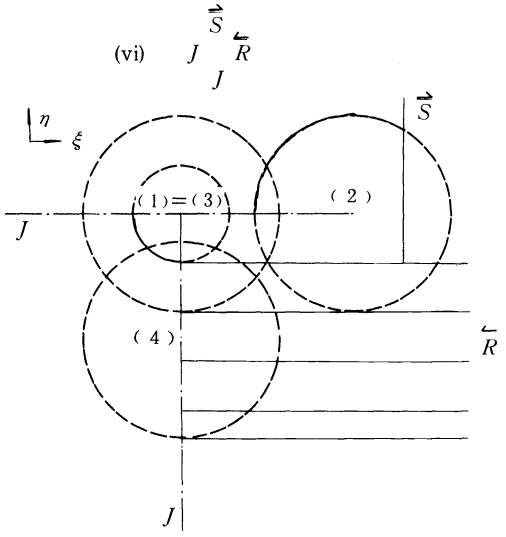


FIG. 7.12

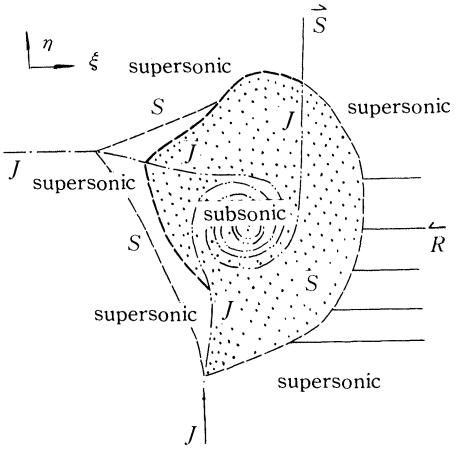


FIG. 7.13

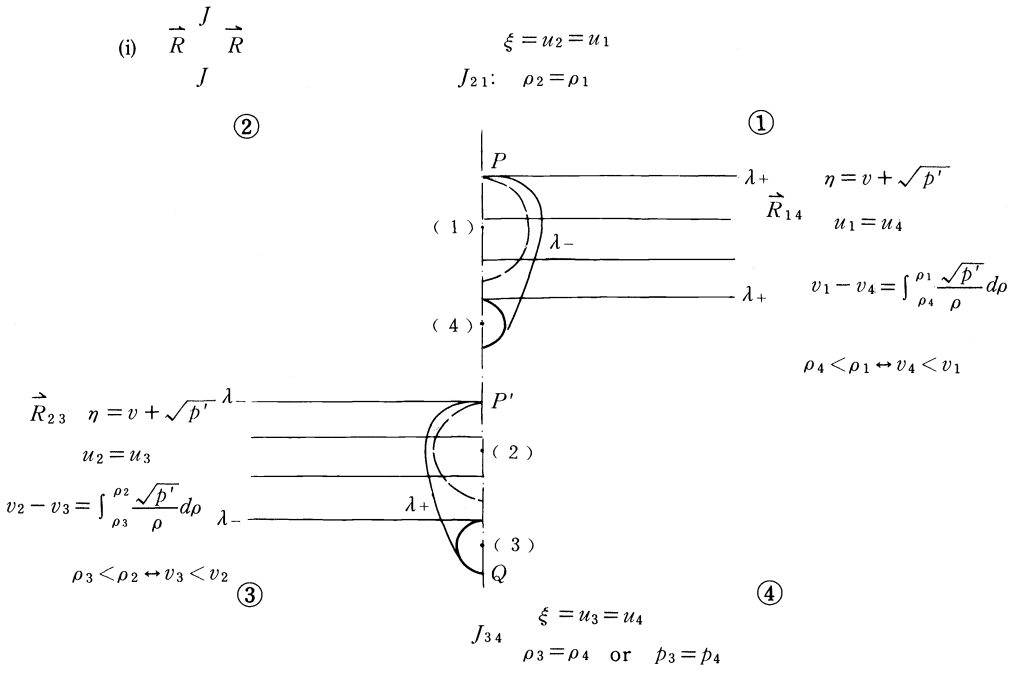


FIG. 7.14

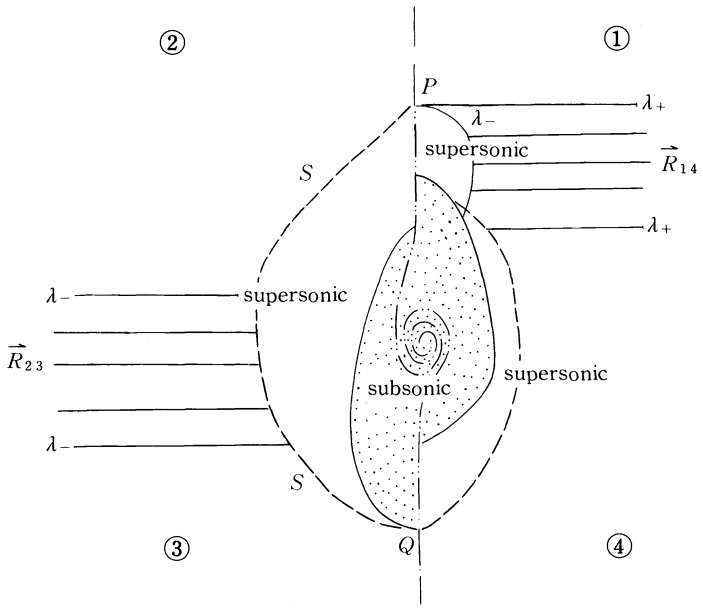


FIG. 7.15

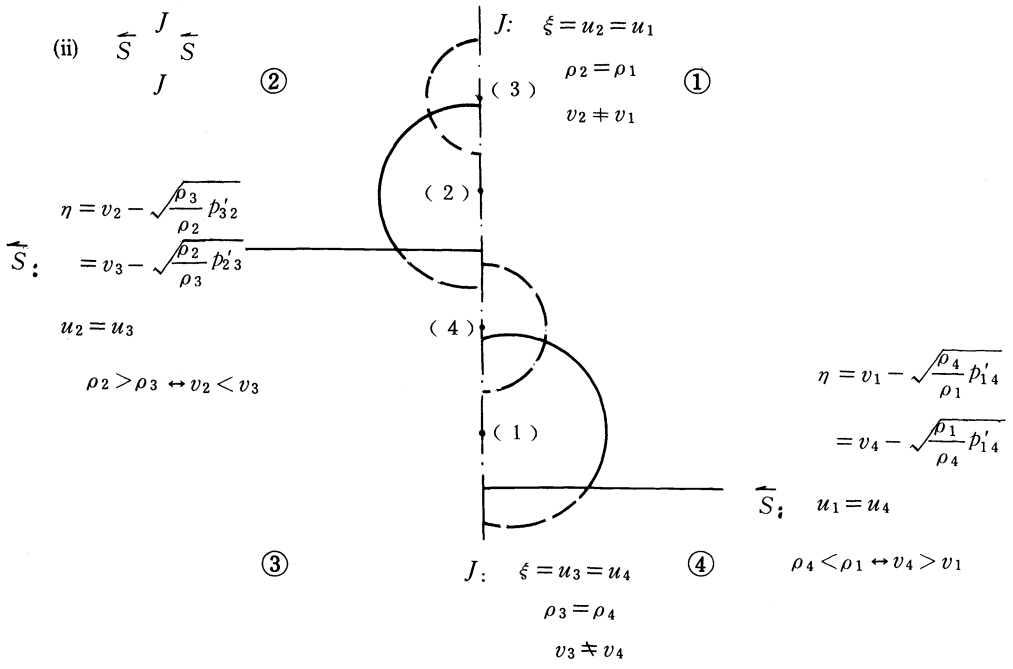


FIG. 7.16

at the side of ④ and be tangent to the λ_+ characteristic there. The shock should be convex with respect to point (4), be tangent to the sonic curve somewhere, and vanish there continuously. There is a subsonic region inside these two shocks and the J's end with spirals in this region (Fig. 7.15).

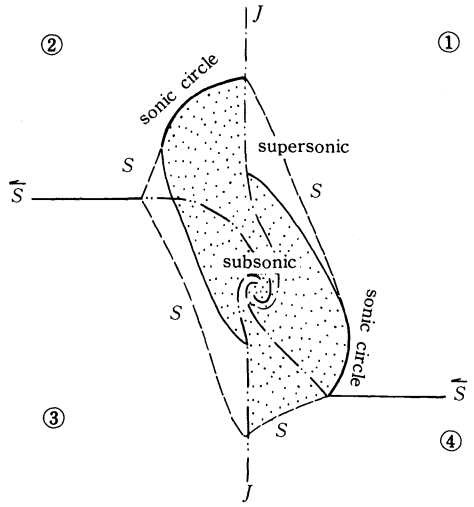


FIG. 7.17

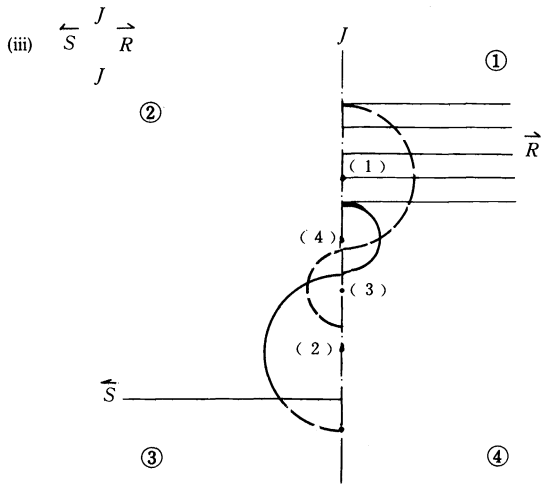


FIG. 7.18

(ii) $\tilde{S}\overset{J}{}_{J}\tilde{S}$ (Fig. 7.16). Divide $\tilde{S}$ along the vertical stream line into two parts, then slip them apart (Fig. 7.16).

We conjecture that the solution is as shown in Fig. 7.17.

(iii) $\tilde{S}\overset{J}{}_{J}\vec{R}$ (Fig. 7.18). Take a half of (i) and half of (ii), and piece them together along the slip line (Fig. 7.18).

We conjecture that the solution is as shown in Fig. 7.19.

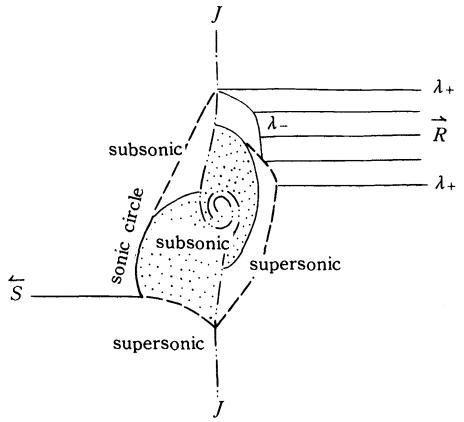


FIG. 7.19

REFERENCES

[1] CHEN-JUN ZHANG, JIN-GEN SUN, JI-WU XIONG, YING-XIN LIAO, SHU-FENG JIA, AND TONG-ZU YANG, *Initial value problem with three constant states as the initial data for aerodynamic system*, Bachelor thesis, Chinese Science and Technology University, Hefei, People's Republic of China, 1963.

[2] TONG CHANG AND YUFA GUO, *A class of initial value problem for systems of gas dynamic equations*, Acta Math. Sinica, 15 (1965), pp. 386-396. (In English.) Chinese Math., 7 (1965), pp. 90-101. (In Chinese.)

[3] SHIQI DING, TUNG CHANG, CHINHUA WANG, LING HSIAO, AND TSAICHUNG LI, *A study of the global solution for quasilinear hyperbolic system of conservation law*, Sci. Sinica, 16 (1973), pp. 317-335.

[4] LING HSIAO AND TONG ZHANG, *Interaction of elementary waves in one dimensional adiabatic flow*, Acta Math. Sinica, 22 (1979), pp. 596-619. (In Chinese.)

[5a] TUNG CHANG AND LING HSIAO, *Riemann problem and discontinuous initial value problem for typical quasilinear hyperbolic system without convexity* (Abstract), Acta Math. Sinica, 20 (1977), pp. 229-231. (In Chinese.)

[5b] LING HSIAO AND TUNG CHANG, *Perturbation on Riemann problem in gas dynamics* (Complete paper), J. Math. Anal. Appl., 79 (1981), pp. 436-460.

[6] TUNG CHANG, TSAICHUNG LI, AND LING HSIAO, *Global solution for a class of initial value problem for a typical quasilinear hyperbolic system without convexity*, Kexue Tongbao, 11 (1975), pp. 506-509. (In Chinese.)

[7] TUNG CHANG AND LING HSIAO, *Riemann problem for one dimensional adiabatic flow without convexity*, Acta Math. Sinica, 20 (1977), pp. 229-231. (In Chinese.)

[8] TONG ZHANG AND YUXI ZHENG, *Two dimensional Riemann problem for a scalar conservation law*, Trans. Amer. Math. Soc., 312 (1989), pp. 589-619.

[9] GUINQIANG CHEN, *Overtaking of shocks in plane steady supersonic isentropic flow*, Master thesis, Institute of Mathematics, Academia Sinica, Beijing, People's Republic of China, 1984.

[10] LING HSIAO AND TONG ZHANG, *Overtaking of shocks belonging to some family in steady plane supersonic flow*, Acta Math. Appl. Sinica, 3 (1980), pp. 343-345. (In Chinese.)

[11] TUNG CHANG AND GUIQIANG CHEN, *Some fundamental concepts for system of two spatial dimensional conservation laws*, Acta Math. Sci., 6 (1986), pp. 463-474.

[12] ———, *Diffraction of planar shock along compressive corner*, Acta Math. Sci., 6 (1986), pp. 241-257.

[13] R. KRASNY, *Desingularization of periodic vortex sheet roll-up*, J. Comput. Phys., 65 (1986), pp. 292-313.

[14] ———, *Computation of vortex sheet roll-up in the Trefftz plane*, J. Fluid Dynamics, 167 (1986), pp. 65-93.

[15] P. COLELLA AND H. GLAZ, *Numerical computation of complex shock reflections in gases*, Ninth Internat. Conference on Numerical Method in Fluid Dynamics, Saclay, France, June 25-29, 1984, Springer Lecture Notes in Physics 218, Springer-Verlag, Berlin, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 631-646, May 1990
© 1990 Society for Industrial and Applied Mathematics
004

ON THE ENSKOG EQUATION WITH LARGE INITIAL DATA*

LEIF ARKERYD†

Abstract. This paper is concerned with the Enskog equation with large initial data in L^1, where the high density factor is constant. As a preliminary step, existence and uniqueness is first studied in full physical space and in a box with periodic boundary conditions under the restriction of bounded velocities, by the use of a priori estimates in the norm $\int (\sup_{0\le t\le T} f(x+tv, v, t)|)\, dx\, dv$. Global existence and uniqueness for small data and unbounded velocities is an easy consequence of this step. The rest of the paper is devoted to the central topic: global existence, regularity, and uniqueness for large initial data in full physical space for the case of unbounded velocities, provided all v-moments are initially finite. Here the more detailed structure of the collision operator is exploited in the a priori estimates.

Key words. Enskog equation, global results, well-posedness, regularity

AMS(MOS) subject classification. 76P05

1. Introduction. The Enskog equation [8] is a quite successful model of transport phenomena in moderately dense gases [9]. This paper studies the initial value problem for the Enskog equation with large L^1 initial data. Previous large data results include global existence in L^1 for one space dimension [5] and two space dimensions [1]. Existence in arbitrary dimension for small data was also obtained in [6], and for large data in [2], [7], and [10]. The averaging argument used there gives no information about, e.g., uniqueness or regularity. The present paper, on the other hand, uses a contraction mapping suitable for extracting detailed information about the solutions.

Let v', v'_* be the velocities after collision of two molecules having precollisional velocities v, v_*, and with v', v'_* also depending on the directional variable u belonging to the unit sphere ∂B. In a suitable coordinate system the Enskog collision operator can be written

$$Q(f,f) = \sigma^2 \int_{\partial B\times R^3} (f'f'_*\chi^- - ff_*\chi^+) S(v, v_*, u)\, dv_*\, du,$$

with

$$(v-v_*, u)_+ = S(v, v_*, u) = \max(0, (v-v_*, u)).$$

The arguments of f', f'_*, f, f_* are (x, v'), $(x-\sigma u, v'_*)$, (x, v), and $(x+\sigma u, v_*)$, where $v' = v - u(v-v_*, u)$ and $v'_* = v_* + u(v-v_*, u)$. The Enskog equation, depending on the variables $(x, v, t) \in R^3\times R^3\times R_+$, is

$$\frac{\partial}{\partial t} f + v\cdot \nabla_x f = Q(f,f). \tag{1.1}$$

The high-density factors $\chi^\pm$ in the original Enskog paper [8] are functions of the local density $\int f(\cdot, v, \cdot)\, dv$ at $(x\pm\sigma u/2, t)$. A later modification that formally satisfies some entropy bound takes $\chi^\pm$ as a function of the local density at (x, t) and $(x\pm\sigma u, t)$. The present paper considers only the simplified case of $\chi^\pm$ = constant, commonly known as the Enskog–Boltzmann equation, and for which a strict proof of an H-theorem is known ([7]; see also (1.4) below). The extension to the case of a variable high-density

* Received by the editors June 27, 1988; accepted for publication (in revised form) April 5, 1989.

† Department of Mathematics, Chalmers University of Technology, Göteborg, Sweden, and University of Göteborg, S-41296 Göteborg, Sweden.

factor must await a strict proof of an H-theorem in that case. For further information on the various forms of the Enskog equation as well as an extensive bibliography, see [3].

This paper starts with a study of existence and uniqueness in the case of bounded velocities by contraction mappings and by conservation properties. For this, mass and moments of second order are assumed to exist initially. The symmetry properties of the collision operator then imply that these quantities are bounded on any bounded time interval. This can be used to control the higher moments of interest in estimates of the type

$$\|f\| \leqq c\|f\| + \delta.$$

Here $0 < c < 1$, and the smallness of δ is controlled independently of the velocity bounds. From such estimates, a priori bounds and local convergence of the approximations with respect to the $\|\cdot\|$-norm are obtained in the unbounded velocity case. Certain uniformity features of the estimates, together with entropy estimates in the unbounded case, lead to the global results. The following regularity discussion depends on similar estimates and bases its control of the moments of the derivatives on the control of the moments of f.

The plan of the paper is as follows. The Introduction continues with a description of the problem. Section 2 is a preliminary study of the case of bounded velocities. Global small-data existence and uniqueness for unbounded velocities is a simple consequence. Section 3 contains some useful estimates independent of the velocity cutoffs. In § 4 the global existence and uniqueness theorem for unbounded velocities is proved. Finally, the regularity properties of those solutions are studied in § 5.

Denote by L_r the space of measurable functions f on $\Omega = R^3 \times R^3$ with $(1 + |v|^r) f(x, v) \in L^1(\Omega)$. Set $|v|_M = \max(|v|, 1)$. The norm $\int_{R^3} |v|_M^r |f(x, v)|\, dx\, dv$ is written $\|\cdot\|_r$, and the positive cone in L_r is L_r^+. Let $L_{r,T}$ be the space of measurable functions on $\Omega \times [0, T]$ with

$$\operatorname*{ess\,sup}_{0 \leqq t \leqq T} |f(x + vt, v, t)| \in L_r,$$

and denote the $\|\cdot\|_r$-norm of this ess sup by $\|\cdot\|_{r,T}$. The use of such norms was introduced into nonlinear kinetic theory by Toscani in a discrete velocity setting [11].

Consider (1.1) for $t > 0$ and regard the left-hand side as a directional derivative $D_t f^{\#}$ along $f^{\#}(x, v, t) = f(x + vt, v, t)$. Take $f_0 = f(0+)$ as initial condition with $f_0 \in L_r^+$ for all $r \in R$. The aim of the paper is to study (1.1) with such initial conditions globally in time also for large f_0.

Starting with the Caflisch paper [4], the different influences of the low and high velocities on the evolution have sometimes been analysed by a corresponding splitting of the density function. Here we will use one such splitting well adapted to the present problem and similar ones in kinetic theory. For that purpose set

$$f_{i0}(x, v) = \min(f_0(x, v), w) \quad \text{for } |x|^2 \leqq w^2, \quad |v|^2 \leqq w^2,$$

and set $f_{i0}(x, v) = 0$ otherwise. Introduce

$$f_i^{\#}(x, v, t) = f_{i0}(x, v), \quad f_e = f - f_i, \quad f_{e0} = f_0 - f_{i0},$$

and $\mathcal{B}^{\#}$ as w times the characteristic function of $\operatorname{supp} f_{i0}$. The equation for f_e is

$$D_t f_e^{\#} = Q(f_i + f_e, f_i + f_e)^{\#}, \qquad t > 0,$$

with initial condition $f_e(0+) = f_{e0}$. Throughout the paper C denotes various constants.

The proofs will use the energy bound

$$\int v^2 f(x, v, t)\, dx\, dv \leqq \int v^2 f_0(x, v)\, dx\, dv, \tag{1.2}$$

a bound of the second x-moment due to J. Polewczak [10]:

$$\int (x - tv)^2 f(x, v, t)\, dx\, dv \leqq \int x^2 f_0(x, v)\, dx\, dv, \tag{1.3}$$

and an entropy bound due to Cercignani [7]:

$$\begin{aligned} &\int f(x, v, t) \log f(x, v, t)\, dx\, dv \\ &\quad \leqq \int f_0(x, v) \log f_0(x, v)\, dx\, dv + \frac{1}{2}\left(\int f_0(x, v)\, dx\, dv\right)^2. \end{aligned} \tag{1.4}$$

(I am grateful to C. Cercignani for giving me access to this result before its publication.) The proofs of (1.2)-(1.4) hold in a formal sense for solutions of (1.1) provided that

$$(1 + v^2 + x^2) f_0, \qquad f_0 \log f_0 \in L^1.$$

Estimates (1.2) and (1.3) can be carried out strictly on various approximations of (1.1) and follow in that way in the limit for the solutions of (1.1) to be constructed below in this paper. The case of (1.4) is treated in the proof of Theorem 4.1. By straightforward computation (1.2)-(1.4) imply

$$\int x^2 f(x, v, t)\, dx\, dv \leqq 2 \int x^2 f_0(x, v)\, dx\, dv + 4t^2 \int v^2 f_0\, dx\, dv, \tag{1.5}$$

$$\begin{aligned} &\int_{f > w} f(x, v, t)\, dx\, dv \\ &\quad \leqq (\log w)^{-1}\left(\int f_0(x, v) \log f_0(x, v)\, dx\, dv + \frac{1}{2}\left(\int f_0(x, v)\, dx\, dv\right)^2\right. \\ &\qquad + (4t^2 + 1) \int v^2 f_0(x, v)\, dx\, dv + \int \exp(-v^2 - x^2)\, dx\, dv \\ &\qquad \left. + 2 \int x^2 f_0(x, v)\, dx\, dv\right). \end{aligned} \tag{1.6}$$

2. Existence and uniqueness in the case of bounded velocities. This section contains a preliminary analysis of the Enskog equation for bounded velocities, i.e., with the collision operator

$$Q^j(f, g) = \sigma^2 \int_{R^3} \int_{\partial B} (f' f'_* - f f_*) W_j S\, dv_*\, du,$$

where $W_j = 1$ if $v^2 + v_*^2 \leqq 2^{2j}$, and $W_j = 0$ otherwise. The relevant Enskog equation in integral form is $A^j f = f$, where

$$(A^j f)^{\#} = f_0 + \int_0^t Q^j(f, f)^{\#}(s)\, ds.$$

In view of § 4, consider for a fixed (large) time interval $[0, T_1]$ in the case where $2^j > 2w$, w being such that the right-hand side in (1.6) with $t = T_1$ is bounded from above by $(512)^{-1}$. That implies

$$\int_{|v|>w} |v| f_0(x, v)\, dx\, dv < 256^{-1}, \qquad \|f_{e0}\|_1 < 128^{-1}. \tag{2.1}$$

(The powers of $\frac{1}{2}$ introduced here and below are chosen for convenience instead of the usual cumbersome statements of the type "there is a small enough constant independent of... such that the following estimate holds," where... usually stands for a couple of lines describing what particular variables and expressions the constant should be independent of.)

Denote by $\mathcal{M}(\delta)$ the set of all measurable subsets $M \subset R^6$ such that for almost every $v \in R^3$ the set M_v of those x for which $(x, v) \in M$ has measure less than δ. Define

$$\mathcal{S}(\delta, F) = \sup_{\mathcal{M}(\delta)} \int_M |F(x, v)|\, dx\, dv.$$

Integrals of the following type will appear in the estimates of the collision operator:

$$\int_0^T ds \int_{R^3} \int_{\partial B} F(y + s(v - v_*) + \sigma u, v_*)(v - v_*, u)_+ \sigma^2 W_j\, dv_*\, du. \tag{2.2}$$

Since the Jacobian of the change of variables $\sigma^2(v - v_*, u)_+\, ds\, du \to dx$ equals 1, they can be bounded by $\mathcal{S}(\delta, F)$ for $\delta = T 2^{j+1} \pi \sigma^2$.

An estimate concerning the function f_i of § 1 will be needed.

LEMMA 2.1.

$$\sigma^2 \int_0^T ds \int |v|_M^r f_i^{\#}(x, v, s) f_i^{\#}(x + s(v - v_*) + \sigma u, v_*, s) S\, dx\, dv\, dv_*\, du$$

$$\leqq \mathcal{S}(2w\pi\sigma^2 T, f_{i0}) \|f_{i0}\|_r.$$

Proof. Use (2.2) and note that $|v - v_*| \leqq 2w$ on the support of the integrand.

LEMMA 2.2. *If $g \in L^+_{r,T}$ and $T > 0$, then*

$$\sigma^2 \int_0^T ds \int |v|_M^r g^{\#}(x, v, s) g^{\#}(x + s(v - v_*) + \sigma u, v_*, s) S\, dx\, dv\, dv_*\, du \leqq \|g\|_{r,T} \|g\|_{0,T}.$$

Proof. The proof follows from

$$\int dv_* \int \sigma^2\, du \int_0^T ds\, g^{\#}(x + s(v - v_*) + \sigma u, v_*, s) S \leqq \|g\|_{0,T}.$$

THEOREM 2.3. *Suppose $f_0(1 + v^2 + x^2) \in L^+_0$, $f_0 \log f_0 \in L_0$. Then, locally in time, the equation $A^j f = f$ has a unique solution $f^j \in L^+_0$ that conserves mass, first v-moments, and energy, and satisfies* (1.3).

Proof. The following approximations will be shown to converge to a solution of $A^j f = f$ on some open neighborhood of $t = 0$. With $f_0 = 0$ and

$$L_j f_n(t) = \sigma^2 \int_0^t ds \int f_n(x + vs + \sigma u, v_*, s)\, W_j S\, dv_*\, du,$$

define inductively for $n \in N$:

$$f_{n+1}^{\#}(t) = f_0 \exp(-L_j f_n(t)) + \int_0^t ds \exp(-L_j f_n(t) + L_j f_n(s))$$

(2.3)

$$\times \int (f'_n f'_{n*})^{\#}(s) \sigma^2 W_j S \, dv_* \, du.$$

Then

$$f_{n+1}^{\#}(t) = f_0 + \int_0^t ds \int (f'_n f'_{n*} - f_{n+1} f_{n*})^{\#}(s) \sigma^2 W_j S \, dv_* \, du.$$

In particular, the perturbation

$$_e f_{n+1}^{\#} = f_{n+1}^{\#} - f_i^{\#}$$

satisfies

$$_e f_{n+1}^{\#}(t) = f_{e0} + \int_0^t ds \int (f'_i f'_{i*} + f'_{ie} f'_{n*} + {}_e f'_n f'_{i*} + {}_e f'_{ne} f'_{n*}$$

$$- f_i f_{i*} - f_{ie} f_{n*} - {}_e f_{n+1} f_{i*} - {}_e f_{n+1} {}_e f_{n*})^{\#}(s) \sigma^2 W_j S \, dv_* \, du = f_{e0} + \mathcal{I}_1 + \cdots + \mathcal{I}_8.$$

Here $\mathcal{I}_1, \cdots, \mathcal{I}_8$ are the eight terms of the collision operator in the given order. This expression gives an $\|\cdot\|_{0,T}$-estimate for f_{n+1}.

By Lemma 2.1 the integral of $|\mathcal{I}_1| + |\mathcal{I}_5|$ is bounded from above by $2\mathcal{S}(2w\pi\sigma^2 T, f_{i0}) \|f_{i0}\|_0$.

The corresponding integral of $|\mathcal{I}_4| + |\mathcal{I}_8|$ is, by Lemma 2.2, bounded by

$$\|_e f_n\|_{0,T} (\|_e f_n\|_{0,T} + \|_e f_{n+1}\|_{0,T}). \tag{2.4}$$

After a change of variables in $\int |\mathcal{I}_2|$, $\int |\mathcal{I}_3|$, each of the integrals of $|\mathcal{I}_2|$, $|\mathcal{I}_3|$, $|\mathcal{I}_6|$ can be bounded by

$$\int_0^T ds \int f_i(x + vs, v, s) |_e f_n(x + vs + \sigma u, v_*, s)| \sigma^2 W_j S \, dv \, dv_* \, dx \, du.$$

As in the proof of Lemma 2.1, this in turn is bounded by

$$\mathcal{S}(2^{j+1} \pi \sigma^2 T, f_{i0}) \|_e f_n\|_{0,T}.$$

In the same way the integral of $|\mathcal{I}_7|$ is bounded by

$$\mathcal{S}(2^{j+1} \pi \sigma^2 T, f_{i0}) \|_e f_{n+1}\|_{0,T}.$$

It follows that

$$\begin{aligned} \|_e f_{n+1}\|_{0,T} \leqq{} & \|f_{e0}\|_0 + 2\mathcal{S}(2w\pi\sigma^2 T, f_{i0}) \|f_{i0}\|_0 \\ & + (\|_e f_n\|_{0,T} + \|_e f_{n+1}\|_{0,T} + 3\mathcal{S}(2^{j+1} \pi \sigma^2 T, f_{i0})) \|_e f_n\|_{0,T} \\ & + \mathcal{S}(2^{j+1} \pi \sigma^2 T, f_{i0}) \|_e f_{n+1}\|_{0,T}. \end{aligned}$$

Choose T so that $\mathcal{S}(2^{j+1} \pi \sigma^2 T, \mathcal{B}) < 16^{-1}$, and $2\mathcal{S}(2w\pi\sigma^2 T, \mathcal{B}) \|\mathcal{B}\|_0 < 128^{-1}$. It follows that $\|_e f_n\|_{0,T} \leqq 16^{-1}$ for $n \in N$. Moreover, for the same value of T,

$$\begin{aligned} \|_e f_{n+1} - {}_e f_{m+1}\|_{0,T} \leqq{} & 3\mathcal{S}(2^{j+1} \pi \sigma^2 T, f_{i0}) \|_e f_n - {}_e f_m\|_{0,T} \\ & + \mathcal{S}(2^{j+1} \pi \sigma^2 T, f_{i0}) \|_e f_{n+1} - {}_e f_{m+1}\|_{0,T} \\ & + (\|_e f_n\|_{0,T} + \|_e f_m\|_{0,T} + \|_e f_{m+1}\|_{0,T}) \|_e f_n - {}_e f_m\|_{0,T} \\ & + \|_e f_n\|_{0,T} \|_e f_{n+1} - {}_e f_{m+1}\|_{0,T} \\ \leqq{} & 8^{-1} \|_e f_{n+1} - {}_e f_{m+1}\|_{0,T} + \tfrac{3}{8} \|_e f_n - {}_e f_m\|_{0,T}. \end{aligned}$$

Hence $({}_e f_n)_0^\infty$ is Cauchy in the $\|\cdot\|_{0,T}$-norm. Denote the limit by f_e^j. It follows that the equation $A^j f = f$ has a unique, nonnegative solution $f^j = f_i + f_e^j$ on $[0, T]$ with the $\|\cdot\|_{0,T}$-norm bounded. The time-integrated gain and loss terms

$$\int_0^T ds \int f^{j\prime} f_*^{j\prime} \sigma^2 W_j S \, dv_* \, du, \qquad \int_0^T ds \int f^j f_*^j \sigma^2 W_j S \, dv_* \, du$$

each belong to L_0, so on $[0, T]$ it follows that the usual change of variable proof can be applied to prove that the mass, v-moments, and energy of f_0 are conserved by f^j, and that (1.3) holds.

Remark. (i) In fact, the proof does not require $v^2 f_0 \in L^1$.

(ii) The proof implies stability. If $(f_{0k})_N$ converges to f_0 in L^1, and if $(\int x^2 f_{0k} \, dx \, dv)_N$ is uniformly bounded, then $(f_k^j(\cdot, t))_N$ converges to $f^j(\cdot, t)$ in L^1 on $[0, T]$.

(iii) The theorem holds with the same proof for a periodic x-domain, and for both types of domain for the symmetrized Enskog equation. (The symmetrized equation uses the whole ∂B as a u-domain in Q and the same v', v_*' for $\pm u$ together with $S(v, v_*, u) = |(v - v_*, u)|$.) Just as in the Boltzmann equation, the H-function $\int f^j(x, v, t) \log f^j(x, v, t) \, dx \, dv$ is nonincreasing for the symmetrized Enskog equation, which in turn implies existence for all $t > 0$.

(iv) If $\|f_0\|_0$ is small enough (e.g., less than 128^{-1}), then a shortened version of the proof above, using only the term (2.4) and the physical space R^3, shows that the full Enskog equation (1.1) without velocity cutoff has a global, unique solution conserving mass and v-moments, and having nonincreasing energy.

3. Bounds for the moments of the collision term. Let f_i (depending on f_0) be as defined in § 1.

LEMMA 3.1. *Suppose that* $w 2^{-k} \leqq \delta \ll \eta \ll 1$, *and that* $g \in L_{r,T}^+$. *Then the following estimate holds*:

$$\sigma^2 \int_0^T dt \int_{|v| \geqq 2^{k+1}} dv |v|_M^r \int (f_i' g_*' + g' f_{i*}')^{\#} S \, dv_* \, du \, dx$$
$$\leqq C\{(1-\delta)^{-r} \eta^2 + (1 + \eta^2 - \delta^2)^{-r/2}\} \|g\|_{r,T}.$$

Here C depends on w but not on f_0, k, or T.

Proof. This is the first of several computations using a splitting of the domain of integration into parts related to the relevant properties of the integrand. The $f_i' g_*'$ and $g' f_{i*}'$-terms can be treated analogously so only the $f_i' g_*'$-term is discussed. Under the present hypotheses,

$$|v'| \leqq w \leqq \delta 2^k < \delta |v|$$

if $(x, v') \in \operatorname{supp} f_i'$. Set $\chi_j(v) = 1$ for $2^j \leqq |v| \leqq 2^{j+1}$, and $\chi_j(v) = 0$ otherwise.

For $j > k$ consider first the part of the domain of integration where $v \in \operatorname{supp} \chi_j$ and $|v_*| < \eta |v|$. From the conservation of first moments and energy for the mapping $(v, v_*) \to (v', v_*')$, it follows that

$$|v - v_*'| < (\eta + \delta)|v|, \qquad |v_*'| > (1 - \delta)|v|.$$

Recall that

$$|f_i(x + tv, v', t)| \leqq w \quad \text{if } |x + t(v - v')| \leqq w, \quad |v'| \leqq w,$$

and $f_i = 0$ otherwise. If

$$G_T(x, v) = \sup_{0 \leqq t \leqq T} |g^{\#}(x, v, t)|,$$

then

$$|g(x-\sigma u+tv, v'_*, t)| \leqq G_T(x-\sigma u+t(v-v'_*), v'_*).$$

This together with a change of variables $x \to x-\sigma u+t(v-v'_*)$, followed by an integration with respect to t, x, v'_*, u, and v' in that order, gives an upper bound for the present part of the integral, namely,

$$C(1-\delta)^{-r}\eta^2\|(\chi_{j-1}+\chi_j+\chi_{j+1})g\|_{r,T}. \tag{3.1}$$

Here η^2 comes from the angular contribution, and C depends linearly on w^5 but does not depend on j or T.

Consider next for $v \in \operatorname{supp} \chi_j$ the remaining part of the integral, i.e., the part with $|v_*| > \eta|v|$. Then

$$|v|^r \leqq (1+\eta^2-\delta^2)^{-r/2}|v'_*|^r,$$

and if $v'_* \in \operatorname{supp} \chi_{j+n}$ with $n \geqq 2$, in addition

$$|v|^r \leqq 2^{-(n-1)r}|v'_*|^r.$$

It follows that this part of the integral can be bounded by

$$\begin{aligned}&(1+\eta^2-\delta^2)^{-r/2}3^{-1}2^3\pi w^5\sigma^2 \int (\chi_j(v'_*)+\chi_{j+1}(v'_*))G_T(x, v'_*)|v'_*|^r\,dx\,dv'_*\,du \\ &+\sum_{n\geqq 2} 2^{-(n-1)r}3^{-1}2^3\pi w^5\sigma^2 \int G_T(x, v'_*)|v'_*|^r\chi_{n+j}(v'_*)\,dx\,dv'_*\,du.\end{aligned} \tag{3.2}$$

Finally, sum (3.1) and (3.2) for $j > k$, and note that

$$\sum_{n\geqq 1} 2^{-nr} \leqq C(1+\eta^2-\delta^2)^{-r/2}.$$

The lemma follows from here.

When we keep the notation of Lemma 3.1, the remaining integral will be estimated next.

LEMMA 3.2.

$$\begin{aligned}&\sigma^2 \int_0^T dt \int_{|v|\leqq 2^{k+1}} dv|v|_M^r \int (f'_i g'_* + g' f'^{*}_i)^{\#} S\,dv_*\,du\,dx \\ &\quad \leqq 2^{r/2+1}\mathscr{S}(T2^{k+3}\pi\sigma^2, F_{iT}|v|_M^r)\|g\|_{r,T} + 2^{-r+1}\|f_i\|_{0,T}\|g\|_{r,T}.\end{aligned}$$

Proof. As in Lemma 3.1 it is enough to consider the $f'_i g'_*$-term. Split the domain of integration into two subdomains, $|v'_*| < 2^{k+2}$ and $|v'_*| \geqq 2^{k+2}$. As in the proof of Lemma 2.1 the first integral gives a contribution bounded by

$$2^{r/2}\mathscr{S}(T2^{k+3}\pi\sigma^2, F_{iT}|v|_M^r)\|g\|_{r,T}.$$

Substitute $|v|_M$ by $|v_*|_M/2$ in the second integral and argue as in the proof of Lemma 2.2 to get the bound

$$2^{-r}\|f_i\|_{0,T}\|g\|_{r,T}.$$

LEMMA 3.3. *For $g, p \in L^+_{r,T}$ and $0 < \delta \ll 1$, set $G_T = \sup_{t\leqq T} g^{\#}(t)$, $P_T = \sup p^{\#}(t)$. The following estimate holds*:

$$\begin{aligned}&\sigma^2 \int_0^T dt \int g' p'_* |v|^r S\,dv_*\,dv\,du\,dx \\ &\quad \leqq (1+(1-\delta)^{-r}+\varepsilon\delta^{-1/2}2^{r/2})(\|g\|_{t,T}\|p\|_{1/2,T}+\|g\|_{1/2,T}\|p\|_{r,T}) \\ &\qquad +2^{r/2}(\mathscr{S}(\mu, G_T)\|p\|_{r,T}+\mathscr{S}(\mu, P_T)\|g\|_{r,T}),\end{aligned}$$

where $\mu = T2\pi\varepsilon^{-2}\sigma^2$.

Proof. The proof uses a splitting of the domain of integration. By the proof of Lemma 2.2 the integral over that part of the domain where

$$|v'| \geqq |v| \quad \text{or} \quad |v'_*| \geqq |v|$$

can be bounded by

$$(\|g\|_{r,T}\|p\|_{0,T} + \|p\|_{r,T}\|g\|_{0,T}).$$

In the same way the integral over the set where

$$|v'| \leqq \delta|v|, \quad (1-\delta^2)^{1/2}|v| \leqq |v'_*| \leqq |v| \quad \text{or} \quad |v'_*| \leqq \delta|v|, \quad (1-\delta^2)^{1/2}|v| \leqq |v'| \leqq |v|$$

can be bounded by

$$(1-\delta)^{-r}(\|g\|_{r,T}\|p\|_{0,T} + \|g\|_{0,T}\|p\|_{r,T}).$$

Also, the integral over the set where

$$\delta|v| < |v'| < |v|, \quad \delta|v| < |v'_*| < |v|, \quad |v| > \varepsilon^{-2}$$

can be bounded by

$$2^{r/2}\varepsilon\delta^{-1/2}(\|g\|_{r,T}\|p\|_{1/2,T} + \|g\|_{1/2,T}\|p\|_{r,T}).$$

Finally, in the remaining domain $|v| \leqq \varepsilon^{-2}$, and $|v'|, |v'_*| \leqq |v|$. Analogously to the proof of Lemma 2.1 this part can be bounded by

$$2^{r/2}(\mathscr{S}(\mu, G_T)\|p\|_{r,T} + \mathscr{S}(\mu, P_T)\|g\|_{r,T}).$$

This completes the proof of the lemma.

LEMMA 3.4. *Suppose* $g, p \in L^+_{r,T}$, *and* $0 < \delta \ll 1$. *Then the following estimates hold*:

$$\begin{aligned} \sigma^2 \int_0^T dt \int g' p'_* (|v|^r_M + |v_*|^r_M) S \, dv_* \, du \, dv \, dx \\ \leqq 2^{r/2+1}\varepsilon^2\|g\|_{r+1,T}\|p\|_{r,T} + 2\|g\|_{0,T}\|p\|_{r,T} \\ + 2(1-\delta)^{-r}\|g\|_{0,T}\|p\|_{r,T} + \varepsilon^{-2r}\delta^{-r}2^{r/2+1}\mathscr{S}(\delta^{-1}\mu, G_T)\|p\|_{r,T}, \end{aligned} \tag{3.3}$$

$$\begin{aligned} \sigma^2 \int_0^T dt \int g p_* (|v|^r_M + |v_*|^r_M) S \, dv_* \, dv \, du \, dx \\ \leqq 2\varepsilon^2\|g\|_{r+1,T}\|p\|_{r,T} + 2\varepsilon^{-2r}\mathscr{S}(\mu, G_T)\|p\|_{r,T} + 2\|g\|_{0,T}\|p\|_{r,T}. \end{aligned} \tag{3.4}$$

Here $\mu = T2\pi\varepsilon^{-2}\sigma^2$.

Proof. First consider (3.3) for $|v|^r_M$. As in the proof of the previous lemmas, the integral over that part of the domain where $|v'| \geqq \varepsilon^{-2}$ or $|v'_*| \geqq |v|$ can be bounded by

$$2^{r/2}\varepsilon^2\|g\|_{r+1,T}\|p\|_{r,T} + \|g\|_{0,T}\|p\|_{r,T}. \tag{3.5}$$

In the rest of the domain $|v'| \leqq \varepsilon^{-2}$ and $|v'_*| \leqq |v|$. If $|v'| < \delta|v|$, then $|v| \geqq |v'_*| > (1-\delta)|v|$, and the integral can be bounded by

$$(1-\delta)^{-r}\|g\|_{0,T}\|p\|_{r,T}. \tag{3.6}$$

If $|v'| > \delta|v|$ and $|v'| \leqq \varepsilon^{-2}$, $|v'_*| \leqq |v|$, then $|v'_*| \leqq \varepsilon^{-2}\delta^{-1}$, and the integral can be bounded by

$$\varepsilon^{-2r}\delta^{-r}2^{r/2+1}\mathscr{S}(\delta^{-1}\mu, G_T)\|p\|_{r,T}. \tag{3.7}$$

Now (3.3) for $|v|^r_M$ follows from (3.5)-(3.7). The proof of (3.3) for $|v_*|^r_M$ and of (3.4) is similar.

4. Global existence and uniqueness in the case of unbounded velocities. In this section the solution f of the full Enskog equation of § 1 will be obtained as a strong L^1-limit of the solutions f^j from Theorem 2.3 when the physical space is R^3. The main result is Theorem 4.1.

THEOREM 4.1. *Suppose that* $(1+|v|)^r f_0$ *belongs to* L_0^+ *for all* $r \geqq 0$, *and that* $x^2 f_0$, $f_0 \log f_0$ *belong to* L_0. *Then the Enskog equation has a unique solution* f *on* R_+ *with* $\|f\|_{r,T} < \infty$ *for* r, $T > 0$, *with mass and first v-moments conserved, and satisfying* (1.2)–(1.4).

Let $[0, T_1]$ be a (large) time interval and choose w, so that (2.1) holds. Set

$$(A_j' f)^{\#}(t) = f_{e0} \exp(-L_j f_i(t)) + \int_0^t \exp(-L_j f_i(t) + L_j f_i(s)) \cdot \left((Q^j(f_i+f, f_i+f)^{\#}(s) + \int \sigma^2 (ff_{i*})^{\#}(s) W_j S \, du \, dv \right) ds.$$

The function $f_e^j = f^j - f_i$ satisfies $f_e^j = A_j' f_e^j$ for $t > 0$.

LEMMA 4.2. *For any* $f \in L_{2,T}$ *with*

$$\sup_{t \leqq T} \int |f(x, v, t)| |v|_M^2 \, dx \, dv \leqq 4 \int f_0(x, v) |v|_M^2 \, dx \, dv$$

the following estimate holds:

$$\|A_j' f\|_{r,T} \leqq \|f_{e0}\|_r + 4\mathcal{S}(2w\pi\sigma^2 T, f_{i0}) \|f_{i0}\|_r + 4\|f\|_{0,T} \|f\|_{r,T} + CT\|f_0\|_2, \qquad 0 \leqq r \leqq 1.$$

Here the constant C *is independent of* j *and depends only on* w.

Proof. The collision operator in A_j' is

$$Q^j(f_i+f, f_i+f) + \int \sigma^2 (ff_{i*}) W_j S \, du \, dv_*$$

$$= \int \{f_i' f_{i*}' - f_i f_{i*} + f_i' f_*' + f' f_{i*}' - f_i f_* + f' f_* - ff_*\} \sigma^2 W_j S \, dv_* \, du \tag{4.1}$$

$$= \mathcal{I}_1 + \cdots + \mathcal{I}_7,$$

where $\mathcal{I}_1, \cdots, \mathcal{I}_7$ are the seven terms of the collision operator in the given order. The exponential factor in A_j' is bounded from above by 1.

By Lemma 2.1 the resulting integral of $|\mathcal{I}_1| + |\mathcal{I}_2|$ is bounded from above by

$$4\mathcal{S}(2w\pi\sigma^2 T, f_{i0}) \|f_{i0}\|_r. \tag{4.2}$$

By Lemma 2.2 the corresponding integral of $|\mathcal{I}_6| + |\mathcal{I}_7|$ is bounded by

$$4\|f\|_{0,T} \|f\|_{r,T}. \tag{4.3}$$

It is easy to see that the integrals of $|\mathcal{I}_3|$, $|\mathcal{I}_4|$, and $|\mathcal{I}_5|$—after a change of variables in the first two—can be bounded by

$$2 \int_0^T dt \int |v|_M^r |v_*|_M^r f_i(x + vt, v, t) |f(x + vt + \sigma u, v_*, t)| \sigma^2 S \, dv \, dv_* \, dx \, du$$

$$\leqq \sigma^2 w^6 2^4 \pi^2 3^{-1} \int |f(x, v_*, t)| \, |v_*|_M^2 \, dx \, dv_* \, dt.$$

By hypothesis this is bounded by

$$\sigma^2 w^6 2^6 \pi^2 3^{-1} T \int f_0(x, v)|v|_M^2 \, dx\, dv = CT\|f_0\|_2, \tag{4.4}$$

where C depends only on w. The lemma follows from (4.2)-(4.4).

Next choose $T<1$ so that

$$CT\|f_0\|_2 < 256^{-1}, \qquad \mathscr{S}(2w\pi\sigma^2 T, \mathscr{B})(4\|\mathscr{B}\|_{r,T}+1) < 256^{-1}.$$

Note that this condition on T depends on w but not on j. It is an easy consequence of Lemma 4.2 that the solution f^j of Theorem 2.3 exists on $[0, T]$ for every j, and that

$$\|f_e^j\|_{1,T} \leqq 32^{-1}.$$

The same results then hold when f_0 is replaced by any function g such that (1.2)-(1.4) are satisfied by $f(\cdot, t) = g$, $t \leqq T_1$.

Set $F_{T'}^j = \sup_{t\leqq T'} f^{j\#}(t)$, $F_{eT'}^j = \sup_{t\leqq T'} |f_e^{j\#}(t)|$.

LEMMA 4.3. *Given $\delta > 0$, there are $\tilde{T} \leqq T$ and $\mu > 0$ depending on w and f_0, but not on j, such that*

$$\mathscr{S}(\mu, F_{eT'}^j) \leqq \delta, \qquad T' \leqq \tilde{T}.$$

Proof. Evidently the lemma holds if $F_{e\tilde{T}}^j$ is replaced by $\mathscr{B}$. Also $F_{eT'}^j \leqq \mathscr{B} + F_{T'}^j$. So it is enough to consider $F_{T'}^j$:

$$F_{T'}^j \leqq f_0 + \int_0^{T'} ds \int (f^{j\prime} f_*^{j\prime})^{\#} \sigma^2 W_j S \, dv_* \, du$$

$$\leqq f_0 + \int_0^{T'} ds \{f_i' f_{i*}' + |f_i' f_{e*}^{j\prime}| + |f_e^{j\prime} f_{i*}'| + |f_e^{j\prime} f_{e*}^{j\prime}|\}^{\#} \sigma^2 W_j S \, dv_* \, du = \mathscr{I}_1' + \cdots + \mathscr{I}_5'.$$

Evidently the j-independent terms $\mathscr{I}_1'$ and $\mathscr{I}_2'$ satisfy the lemma. By the proof of (4.4), and with $\tilde{T}$ small enough, the lemma holds for $\mathscr{I}_3' + \mathscr{I}_4'$. Finally, for $\mu = 2^{k+2}\pi\sigma^2\tilde{T}$,

$$\int_\Omega \mathscr{I}_5' \, dx\, dv = \int_\Omega \int_0^{T'} \int |f_e^{j\prime} f_{e*}^{j\prime}|^{\#} \sigma^2 W_j S \, dx\, dv\, dv_*\, du\, ds$$

$$\leqq 2^{-k+2} \|f_e^j\|_{1,T'} \|f_e^j\|_{0,T'} + \int_0^{T'} \int |f_e^j f_{e*}^j| \sigma^2 W_{2k} S \, dx\, dv\, dv_*\, du\, ds$$

$$\leqq 2^{-k+2} \|f_e^j\|_{1,T'} \|f_e^j\|_{0,T} + \sup_{\mathscr{M}(\mu)} \int_M |F_{eT'}^j| \, dx\, dv \|f_e^j\|_{0,T'}.$$

Integrating the inequality shown above for $F_{eT'}^j$ over M, taking the supremum over all $M \in \mathscr{M}(\mu)$, and moving the last term to the left-hand side gives

$$\sup_{\mathscr{M}(\mu)} 2^{-1} \int_M F_{eT'}^j \, dx\, dv \leqq \sup_{\mathscr{M}(\mu)} \int_M f_{i0} \, dx\, dv$$

$$+ \sup_{\mathscr{M}(\mu)} \int_M (\mathscr{I}_1' + \mathscr{I}_2' + \mathscr{I}_3' + \mathscr{I}_4') \, dx\, dv + 2^{-k-8}. \tag{4.5}$$

For $\delta > 0$ given, μ, k, and $\tilde{T}$ can be so chosen that for $T' \leqq \tilde{T}$, each of the six terms to the right in (4.5) can be bounded by $\delta/12$. This implies the statement of the lemma.

Remark. If $0 \leqq t \leqq T_1$ and (1.2)-(1.4) are satisfied by $f(\cdot, t) = g$, then f_0 can be replaced by g in Lemma 4.3, and $\tilde{T}$ can be chosen independently of g and t for $0 \leqq t \leqq T_1$.

LEMMA 4.4. *Let $f \in L_{r,T}$ and set $F_T = \sup_{t \leqq T} f^{\#}(t)$. The following estimate holds:*

$$\begin{aligned}\|A_j' f\|_{r,T} \leqq \|f_{e0}\|_r &+ 2^{r/2+2}\mathscr{S}(2w\pi\sigma^2 T, \mathscr{B})\|f_{i0}\|_r \\ &+ C\{(1-\delta)^{-r}\eta^2 + (1+\eta^2-\delta^2)^{-r/2} + 2^{r/2+2}\mathscr{S}(2^{k+3}\pi\sigma^2 T, \mathscr{B}|v|_M^r) \\ &\quad + 2^{-r}\tilde{w} + 2^{r/2+1}\mathscr{S}(2\pi\varepsilon^{-2}\sigma^2 T, F_T)\}\|f\|_{r,T} \\ &+ (3 + 2(1-\delta)^{-r} + \varepsilon\delta^{-1/2}2^{r/2+1})\|f\|_{1/2,T}\|f\|_{r,T}, \qquad r \geqq 1.\end{aligned}$$

Here $w2^{-k} \leqq \delta \ll \eta \ll 1$, $\tilde{w} = w^7\pi^2\sigma^2 9^{-1}2^5$, and the constant C depends only on w.

Proof. As in the proof of Lemma 4.2 the exponential factor in A_j' is bounded by 1, and the integral of $|\mathscr{I}_1| + |\mathscr{I}_2|$ is bounded by

$$2^{r/2+2}\mathscr{S}(2w\pi\sigma^2 T, f_{i0})\|f_{i0}\|_r. \tag{4.6}$$

The integral of $|\mathscr{I}_3| + |\mathscr{I}_4|$ can, by Lemmas 3.1 and 3.2, be bounded by

$$C\{(1-\delta)^{-r}\eta^2 + (1+\eta^2-\delta^2)^{-r/2} + 2^{r/2+1}\mathscr{S}(2^{k+3}\pi\sigma^2 T, f_{i0}|v|_M^r) + 2^{-r}\tilde{w}\}\|f\|_{r,T}. \tag{4.7}$$

Here C depends only on w, and $w2^{-k} \leqq \delta \ll \eta \ll 1$.

The domain of integration for $|\mathscr{I}_5|$ can be split into two parts, $|v_*| \leqq 2^k$ and $|v_*| > 2^k$, and the integral bounded by

$$\{\mathscr{S}(2^{k+3}\pi\sigma^2 T, f_{i0}|v|_M^r) + w^r 2^{-rk}\tilde{w}\}\|f\|_{r,T}. \tag{4.8}$$

By Lemma 3.3 the integral of $|\mathscr{I}_6|$ can be bounded by

$$\{(2 + 2(1-\delta)^{-r} + \varepsilon\delta^{-1/2}2^{r/2+1})\|f\|_{1/2,T} + 2^{r/2+1}\mathscr{S}(\mu, F_T)\}\|f\|_{r,T}, \tag{4.9}$$

where $\mu = T2\pi\varepsilon^{-2}\sigma^2$. Finally, the integral of $|\mathscr{I}_7|$ can be bounded by

$$\|f\|_{r,T}\|f\|_{0,T}. \tag{4.10}$$

The lemma follows from (4.6)-(4.10), since $f_{i0} \leqq \mathscr{B}$, and $w2^{-k} < 2^{-1}$.

A final lemma, which follows next, requires some further w-dependent conditions on T, which are needed for estimates of $A_j' f_e^j$ in the equation $f_e^j = A_j' f_e^j$ using Lemma 4.4. Namely, in Lemma 4.4 choose in turn η suitably small, $r = r_0$ suitably large, δ suitably small, k suitably large, ε suitably small, and $\tilde{T}$ from Lemma 4.3 (all choices only depending on w), so that for $T \leqq \tilde{T}$

$$0.5\|f_e^j\|_{r_0,T} \leqq \|f_{e0}\|_{r_0} + 256^{-1},$$

with r_0 and $\tilde{T}$ depending only on w and not on j. Moreover, r_0 and $\tilde{T}$ can be chosen so that also

$$\|f_e^j\|_{r_0-1,T} \leqq 2(\|f_{e0}\|_{r_0-1} + 256^{-1}),$$

with r_0 and $\tilde{T}$ depending only on w. In the same way, note that for $r > r_0$, $\tilde{T}_r$ depending only on w and r can be chosen so that for $T_r \leqq \tilde{T}_r$

$$\begin{aligned}\|f_e^j\|_{r-1,T_r} &\leqq 2(\|f_{e0}\|_{r-1} + 256^{-1}), \\ \|f_e^j\|_{r,T_r} &\leqq 2(\|f_{e0}\|_r + 256^{-1}).\end{aligned} \tag{4.11}$$

As above, $\tilde{T}$ and $\tilde{T}_r$ can be chosen independent of t and of $f_0 = g$, when (1.2)-(1.4) are satisfied by $f(\cdot, t) = g$, $0 \leqq t \leqq T_1$.

LEMMA 4.5. $\lim_{j\to\infty} \sup_{j' \geqq j} \|A_{j'}' f_e^{j'} - A_j' f_e^{j'}\|_{r-1,T_r} = 0$.

Proof. Take $2^{j-1} \geqq w$ and $j' > j$. Then

$$|\exp(-L_{j'} f_i(t)) - \exp(-L_j f_i(t))|$$

equals zero, if $|v|^2 < 2^{2j} - w^2$. Otherwise it is bounded by 1. It follows that

$$\|A'_{j'}f_e^{j'} - A'_j f_e^{j'}\|_{r-1,T_r}$$

$$\leqq \int_{|v|>2^{j-1}} f_{e0}(x,v)|v|^{r-1}\,dx\,dv + \int_{|v|>2^{j-1}} dv \int_0^{T_r} ds \int |v|^{r-1}|Q^{j'}(f^{j'}, f^{j'})^{\#}(s)|$$

$$+ \sigma^2 \int (f_e^{j'} f_{i*})^{\#}(s)\, W_{j'} S\,du\,dv_*\,dv\,dx + \int_0^{T_r} ds \int (|f'_i f_{e*}^{j'\prime}| + |f_e^{j'\prime} f'_{i*}|$$

$$+ |f_i f_{e*}^{j'}| + |f_e^{j'\prime} f_{e*}^{j'\prime}| + |f_e^{j'} f_{e*}^{j'}|)|v|_M^{r-1}(W_{j'} - W_j) S\sigma^2\,du\,dv\,dv_*\,dx.$$

Evidently the limit when $j \to \infty$ is zero for the first term and the $f_i f_{i*}$-terms. The other terms are bounded by

$$2^{-j+r/2+2}\|f_e^{j'}\|_{r,T_r}(\|f_i\|_{0,T_r} + \|f_e^{j'}\|_{0,T_r}) \leqq 2^{-j+r/2+3}(\|f_{e0}\|_r + 256^{-1})(2\|f_0\|_0 + 32^{-1}).$$

This tends to zero when $j \to \infty$.

Proof of Theorem 4.1. The values of r_0 and $\tilde{T}_r$ given before Lemma 4.5 will be used. First the Cauchy property of the sequence (f_e^j) will be proved in $L_{r-1,\tilde{T}}$ for some $\tilde{T}$ with $0 < \tilde{T} \leqq \tilde{T}_r$ when $r = r_0$. Consider for $j' > j$

$$\begin{aligned} \|f_e^{j'} - f_e^j\|_{r-1,T'} &= \|A'_{j'}f_e^{j'} - A'_j f_e^j\|_{r-1,T'} \\ &\leqq \|A'_{j'}f_e^{j'} - A'_j f_e^{j'}\|_{r-1,T'} + \|A'_j f_e^{j'} - A'_j f_e^j\|_{r-1,T'}. \end{aligned} \tag{4.12}$$

It is enough to estimate the right side of (4.12) by a term equal to a small multiple (less than 1) of $\|f_e^{j'} - f_e^j\|_{r-1,T'}$ plus a term tending to zero when $j \to \infty$.

By the same reasoning as in the proof of Lemma 4.4, the term $\|A'_j f_e^{j'} - A'_j f_e^j\|_{r-1,T'}$ can be bounded by

(4.13)

$$C\{(1-\delta)^{-r}\eta^2 + (1+\eta^2-\delta^2)^{-r/2} + 2^{r/2+2}\mathcal{S}(2^{k+3}\pi\sigma^2 T, \mathcal{B}|v|_M^r) + 2^{-r+1}\tilde{w}\}\|f_e^{j'} - f_e^j\|_{r-1,T'}$$

$$+ \left\| \sigma^2 \int_0^t ds \int \{|f_e^{j'\prime} f_{e*}^{j'\prime} - f_e^{j\prime} f_{e*}^{j\prime}| + |f_e^{j'} f_{e*}^{j'} - f_e^j f_{e*}^j|\}^{\#} W_{j'} S\,du\,dv_* \right\|_{r-1,T'}.$$

Here the constants k, η, and δ have the values given before Lemma 4.5. By Lemma 3.4 the last term in (4.13) can be bounded by

$$\begin{aligned} &\|f_e^{j'} - f_e^j\|_{r-1,T'}\{2^{r/2+2}\varepsilon^2(\|f_e^{j'}\|_{r,T'} + \|f_e^j\|_{r,T'}) \\ &\quad + (4 + 2(1-\delta)^{1-r})(\|f_e^{j'}\|_{0,T'} + \|f_e^j\|_{0,T'}) \\ &\quad + (\mathcal{S}(\delta^{-1}\mu, F_{eT'}^{j'}) + \mathcal{S}(\delta^{-1}\mu, F_{eT'}^j))2^{r/2+2}\varepsilon^{-2r}\delta^{-r}\}, \end{aligned} \tag{4.14}$$

where $\mu = T'2\pi\varepsilon^{-2}\sigma^2$. The factor $2^{r/2+2}\varepsilon^2(\|f_e^{j'}\|_{r,T'} + \|f_e^j\|_{r,T'})$ can be made arbitrarily small by a suitable choice of ε.

Note that if we construct a solution f successively on small subintervals with the same length $\tilde{T}_{r_0}$ and in $[0, T_1]$, then (4.11) gives a uniform upper bound when $T \leqq T_1$ of $\|f\|_{r,T}$, and of $\|f_e^j\|_{r,T_r}$ when f^j has initial value $f(T)$. This upper bound can be inserted instead of $\|f_e^{j'}\|_{r,T'}$, $\|f_e^j\|_{r,T'}$, and ε can be chosen once and for all on $[0, T_1]$. Then by Lemma 4.3 for a small enough $T' = \tilde{T} > 0$ depending only on w, the factor

$$\varepsilon^{-2r}\delta^{-r}2^{r/2+2}(\mathcal{S}(\delta^{-1}\mu, F_{e\tilde{T}}^{j'}) + \mathcal{S}(\delta^{-1}\mu, F_{e\tilde{T}}^j))$$

can be made suitably small, again uniformly on $[0, T_1]$ with respect to f^j with initial value $f(T)$, $0 \leqq T \leqq T_1$. Therefore the estimates (4.13), (4.14) imply

$$\|A'_{j'}f_e^{j'} - A'_j f_e^j\|_{r-1,\tilde{T}} \leqq \|f_e^{j'} - f_e^j\|_{r-1,\tilde{T}}/2,$$

which inserted into (4.12) gives

$$\|f_e^{j'}-f_e^j\|_{r-1,\tilde{T}} \leqq \|f_e^{j'}-f_e^j\|_{r-1,\tilde{T}}/2+\|A_j^l f_e^{j'}-A_j^l f_e^{j'}\|_{r-1,\tilde{T}}.$$

By Lemma 4.5 this implies that the sequence (f_e^j) is Cauchy in the $\|\cdot\|_{r-1}$-norm for $t \leqq \tilde{T}$. It follows that the limit f_e has $(1+|v|^{r-1})f_e(x, v, t)$ in $L^1(\Omega)$ for $t \leqq \tilde{T}$, and that the sum $f=f_i+f_e$ is the unique solution in L^+_{r-1} of

$$(f_i+f_e)^{\#}(t)=f_0+\int_0^t Q(f_i+f_e, f_i+f_e)^{\#}(s)\,ds, \qquad t \leqq \tilde{T}.$$

Moreover, f conserves mass and v-moments and satisfies (1.2) and (1.3) for $t \leqq \tilde{T}=\tilde{T}_{r_0}$.

Next, given $r \geqq r_0$, $f(\cdot, t)$ has finite $\|\cdot\|_r$-norm for $0 \leqq t \leqq \tilde{T}_r$ by (4.11). It follows from the local part of the regularity proof below in Theorem 5.2 that locally around $t=0$ and for r large enough, the solution conserves the x-differentiability of the initial value with respect to the r-norm. By this x-differentiability, the same type of argument implies that the solution conserves the v-differentiability of the initial value locally around $t=0$. From here the formal argument in the proof of the entropy bound (1.4) holds in a strict sense for sufficiently smooth initial values. The initial value f_0 of the present theorem can be approximated by smooth ones, so the continuous dependence of the solution on the initial value in any $\|\cdot\|_{r,T}$-norm then implies (1.4) for the solution f with initial value f_0 on $[0, \tilde{T}_r]$. From here, since $f(\cdot, \tilde{T}_r)$ satisfies (1.2)–(1.4), the same estimate now applies on $[\tilde{T}_r, 2\tilde{T}_r]$, and by induction on $[0, \tilde{T}_{r_0}]$. Also, $\|f(\cdot, \tilde{T}_{r_0})\|_r$ is finite for all r, and $f(\cdot, \tilde{T}_{r_0})$ satisfies (1.2)–(1.4). The whole construction can now be repeated on $[\tilde{T}_{r_0}, 2\tilde{T}_{r_0}]$, and by induction on $[0, T_1]$. Hence there exists a unique solution with all the desired properties on $[0, T_1]$. But T_1 is arbitrary, so the solution exists on R_+. This completes the proof of the theorem.

5. Regularity. The solutions of the Enskog equation shown above retain the regularity properties of the initial value f_0.

THEOREM 5.1. *If $f \in \bigcap_{r\geqq 0} C_0(R_+, L_r)$, with $\|f\|_{r,T}<\infty$ for r, $T>0$, is the solution of the Enskog equation with initial value f_0, where $D^\alpha f_0 \in \bigcap_{r\geqq 0} L_r$ for $|\alpha| \leqq k$, then $D^\alpha f \in \bigcap_{r\geqq 0} C_0(R_+, L_r)$ for $|\alpha| \leqq k$.*

To prove this final result the following technical lemma is needed.

LEMMA 5.2. *Suppose that $f \in \bigcap_{r\geqq 0} C_0(R_+, L_r)$, with $\|f\|_{r,T}<\infty$ for r, $T>0$. Take $T_1>0$ and define*

$$\mathscr{L}g(x, v, t)=\int g(x+tv+\sigma u, v_*, t)S\,dv_*\,du.$$

There is $r_1>r_0$, such that for $r \geqq r_1$, $0 \leqq t_0 \leqq T_1$, and $g \in \bigcap_{r\geqq 0} C_0([0, T_1], L_r)$, with $\|g\|_{r,T}<\infty$ for r, $T>0$, the mapping

$$g \to -\int_{t_0}^t \mathscr{L}g(x, v, \tau)\,d\tau f^{\#}(x, v, t_0)\exp\left(-\int_{t_0}^t \mathscr{L}f(x, v, \tau)\,d\tau\right)-\int_{t_0}^t ds\int_s^t \mathscr{L}g(x, v, \tau)\,d\tau$$

$$\cdot \exp\left(-\int_s^t \mathscr{L}f(x, v, \tau)\,d\tau\right)\int (f'f'_*)^{\#}S\sigma^2\,du\,dv_* \tag{5.1}$$

$$+\int_{t_0}^t ds\exp\left(-\int_s^t \mathscr{L}f(x, v, \tau)\,d\tau\right)\int (f'g'_*+g'f'_*)^{\#}S\sigma^2\,du\,dv_*$$

is strictly contracting (constant of contraction less than 1) in the norm $\int_\Omega \sup_{A_{t_0}}|g^{\#}(x, v, s)|\,|v|_M^r\,dx\,dv$, for some $T_{t_0}>0$. Here

$$A_{t_0}=[t_0-T_{t_0}, t_0+T_{t_0}]\cap[0, T_1].$$

Proof. Set $G_T(\cdot) = \sup_{[0,T]} |g^{\#}(\cdot, t)|$, $F_T = \sup_{[0,T]} f^{\#}(\cdot, t)$. The following estimates hold for the various terms of the mapping (5.1) with $t_0 = 0$. For

$$\int_0^T ds \int |v|_M^r f_0(x, v)|g^{\#}(x + s(v - v_*) + \sigma u, v_*, t)|\sigma^2 S\, dx\, dv\, dv_*\, du$$

estimate separately the part of the domain of integration where $\max(|v|, |v_*|) \leqq 2^k$, and the rest of the domain. With $\mu = 2^{k+1}\sigma^2\pi T$ this gives an upper bound

$$2^{-k+1}(\|g\|_{1,T}\|f_0\|_r + \|g\|_{0,T}\|f_0\|_{r+1}) + \mathcal{S}(\mu, f_0|v|_M^r)\|g\|_{0,T}. \tag{5.2}$$

For

$$\int_0^T ds \int (|v|_M^r + |v_*|_M^r) f^{\#}(x, v', s) \cdot |g^{\#}(x + s(v' - v'_*) - \sigma u, v'_*, s)|\sigma^2 S\, dx\, dv\, dv_*\, du, \tag{5.3}$$

consider first the $|v|_M^r$-term. The integral over the subdomain where $\max(|v'|, |v'_*|) \leqq 2^k$ can be bounded from above by

$$2^{r/2}\mathcal{S}(\mu, F_T|v|_M^r)\|g\|_{r,T}. \tag{5.4}$$

The subdomain where $|v'| > 2^k$ gives the upper bound

$$2^{r/2+2-k}\|f\|_{r+1,T}\|g\|_{r,T}. \tag{5.5}$$

Finally, as in the proof of Lemmas 3.1 and 3.2, the subdomain where $|v'| \leqq 2^k$, $|v'_*| > 2^k$ gives the upper bound

$$\begin{aligned} &C((1-\delta)^{-r}\sigma^2 + (1 + \eta^2 - \delta^2)^{-r/2})\|f\|_{0,T}\|g\|_{r,T} \\ &\quad + (2^{r/2}\mathcal{S}(2^{\nu+3}\sigma^2\pi T, F_T|v|_M^r) + 2^{-r+1}\|f\|_{0,T})\|g\|_{r,T}. \end{aligned} \tag{5.6}$$

Here $2^{k-\nu} \leqq \delta \ll \eta \ll 1$. The following upper bound for the same term can be obtained similarly:

$$\begin{aligned} &C(1-\delta)^{-r}\eta^2\|f\|_{0,T}\|g\|_{r,T} + C(1 + \eta^2 - \delta^2)^{-r/2}2^{-\nu}\|g\|_{r+1,T}\|f\|_{0,T} \\ &\quad + 2^{r/2}\mathcal{S}(2^{\nu+3}\sigma^2\pi T, F_T|v|_M^r)\|g\|_{r,T}. \end{aligned} \tag{5.7}$$

The $|v_*|_M^r$-term gives the same contributions, so an upper bound for (5.3) is

$$\begin{aligned} &\{2^{r/2}\mathcal{S}(\mu, F_T|v|_M^r) + 2^{r/2+2-k}\|f\|_{r+1,T} + C((1-\delta)^{-r}\eta^2 + (1 + \eta^2 - \delta^2)^{-r/2})\|f\|_{0,T} \\ &\quad + (2^{r/2}\mathcal{S}(2^{\nu+1}\sigma^2\pi T, F_T|v|_M^r) + 2^{-r+1}\|f\|_{0,T})\}\|g\|_{r,T}. \end{aligned} \tag{5.8}$$

Next consider the term

$$\begin{aligned} &\int_0^T ds \int_s^T \int g^{\#}(x + \tau(v - \bar{v}_*) + \sigma\bar{u}, \bar{v}_*, \tau)\sigma^2 S\, d\tau\, d\bar{u}\, d\bar{v}_* \\ &\quad \cdot \int f^{\#}(x + s(v - v'), v', s) f^{\#}(x + s(v - v'_*) - \sigma u, v'_*, s)\sigma^2 S|v|_M^r\, dx\, dv\, dv_*\, du. \end{aligned} \tag{5.9}$$

Here

$$\int_s^T \int g^{\#}(x + \tau(v - \bar{v}_*) + \sigma\bar{u}, \bar{v}_*, \tau)\sigma^2 S\, d\tau\, d\bar{u}\, d\bar{v}_* \leqq \|g\|_{0,T}. \tag{5.10}$$

Arguing as in the proof of (5.8) but using (5.7) instead of (5.6), we get the following bound for (5.9):

$$\|g\|_{0,T}\{2^{r/2}\mathscr{S}(\mu, F_T|v|_M^r)\|f\|_{r,T}+2^{r/2+2-k}\|f\|_{r+1,T}\|f\|_{r,T}$$

$$+C(1-\delta)^{-r}\eta^2\|f\|_{0,T}\|f\|_{r,T}+C(1+\eta^2-\delta^2)^{-r/2}2^{-\nu}\|f\|_{r+1,T}\|f\|_{0,T} \tag{5.11}$$

$$+2^{r/2}\mathscr{S}(2^{\nu+3}\sigma^2\pi T, F_T|v|_M^r)\|f\|_{r,T}\}.$$

The terms in (5.1) can be bounded by (5.2), (5.8), or (5.11). In (5.2) with r given choosing k large and then T sufficiently small, the coefficient of $\|g\|_{1,T}$ can be made less than $\frac{1}{4}$. Note that this also can be made to hold if $\|f\|_{r+1,T}$ and $\|f_0\|_r$ are replaced by $\|f\|_{r+1,T_1}$ and $\|g\|_{0,T}$ and $\|g\|_{1,T}$ is replaced by $\|g\|_{r,T}$. Moreover, $\mathscr{S}(\mu, f_0|v|_M^r)$ can be replaced by $\mathscr{S}(\mu, f^{\#}(t_0)|v|_M^r)$ for $T=T_{t_0}$ sufficiently small.

In (5.8) choose, in turn, η suitably small, $r\geqq r_1$ with r_1 suitably large, δ suitably small, k suitably large, ν suitably large (for (5.6)), and T suitably small to make the coefficient of $\|g\|_{r,T}$ less than $\frac{1}{4}$. Again this can be made to hold if $\|f\|_{r+1,T}$ and $\|f\|_{0,T}$ are replaced by $\|f\|_{r+1,T_1}$, $\|f\|_{0,T_1}$, and $\mathscr{S}(\cdots F_T\cdots)$ by $\mathscr{S}(\cdots P_{t_0}\cdots)$, where $P_{t_0}=\sup_{A_{t_0}} f^{\#}(t)$ for a small enough T_{t_0}. In the same way, given r, T_{t_0} can be chosen so small that in (5.11) the coefficient of $\|g\|_{0,T}$ can be made less than $\frac{1}{4}$. From here the lemma follows.

Proof of Theorem 5.1. Given $T_1>0$ it is enough to find $r_k>0$ so that the theorem holds on $[0, T_1]$ for $r\geqq r_k$ and $|\alpha|\leqq k$. Consider first the case $k=1$, and take $D^\alpha=\partial_{x_1}$, $D^\alpha=\partial_{x_2}$, or $D^\alpha=\partial_{x_3}$. It follows by the contraction properties of Lemma 5.2 that, for some $T_r>0$ depending on r, the equation

$$g^{\#}(x, v, t)$$

$$=\left(D^\alpha f_0-f_0\int_0^t \mathscr{L}g(x, v, \tau)\,d\tau\right)\exp\left(-\int_0^t \mathscr{L}f(x, v, \tau)\,d\tau\right)$$

$$+\int_0^t ds\exp\left(-\int_s^t \mathscr{L}f(x, v, \tau)\,d\tau\right)\left\{\int (f'g'_*+g'f'_*)^{\#}\sigma^2 S\,dx\,dv\,dv_*\,du\right\} \tag{5.12}$$

$$-\int_0^t ds\left(\int_s^t \mathscr{L}g(x, v, \tau)\,d\tau\right)\exp\left(-\int_s^t \mathscr{L}f(x, v, \tau)\,d\tau\right)f'f'_*\sigma^2 S\,dx\,dv\,dv_*\,du$$

has a unique solution with $\|g\|_{r,T_r}<\infty$ belonging to $C_0([0, T_1], L_r)$ for $r\geqq r_1$.

Similarly, the difference quotient $\Delta^\alpha f/\Delta x_\alpha$, which solves an equation related to (5.12), can be shown to converge to g in the $\|\cdot\|_{r,T}$-norm when $\Delta x_\alpha\to 0$. Hence $D^\alpha f\in C_0([0, T_1], L_r)$ and $\|D^\alpha f\|_{r,T_r}<\infty$. From here, by the same argument, kth-order differentiability holds locally around $t=0$. In this part of the proof, given r, only the bounds on $\|f\|_{r+1,T_r}$ were used to obtain the contracting properties of Lemma 5.2. For that reason this part of the proof can be used as part of the proof of Theorem 4.5.

Next use Lemma 5.1 in its full force and repeat the argument above with $T_r=T_1$ and $k=1$. It follows that $D^\alpha f\in C_0([0, T_1], L_r)$ for $|\alpha|=1$. Since $T_1>0$ is arbitrary the theorem is proved for $k=1$. Provided the theorem holds for derivatives of order less than or equal to $k-1$, repeating the argument gives the theorem for kth-order derivatives.

Acknowledgment. The author thanks J. Polewczak for pointing out a mistake in an earlier version of the paper.

REFERENCES

[1] L. ARKERYD, *On the Enskog equation in two space variables*, Transport Theory Statist. Phys., 15 (1986), pp. 673-691.

[2] L. ARKERYD AND C. CERCIGNANI, *On the convergence of solutions of the Enskog equation to solutions of the Boltzmann equation*, Comm. in PDE, 14 (1989), pp. 1071-1090.

[3] N. BELLOMO AND M. LACHOWICZ, *Kinetic equations for dense gases*, Internat. J. Modern Phys. B, 1 (1987), pp. 1193-1205.

[4] R. CAFLISCH, *The fluid dynamic limit of the nonlinear Boltzmann equation*, Comm. Pure Appl. Math., 33 (1980), pp. 651-666.

[5] C. CERCIGNANI, *Existence and global solutions for the space inhomogeneous Enskog equation*, Transport Theory Statist. Phys., 16 (1987), pp. 214-222.

[6] ———, *Small data existence for the Enskog equation in L^1*, J. Statist. Phys. 51 (1988), pp. 291-297.

[7] C. CERCIGNANI AND L. ARKERYD, *Global existence in L^1 for the Enskog equation*, preprint, 1989.

[8] D. ENSKOG, *Kinetische Theorie der Wärmeleitung, Reibung und Selbstdiffusion in gewissen verdichteten Gasen und Flüssigkeiten*, Kungl. Svenska Vetenskapsakad. Handl., 63 (1922), pp. 3-44.

[9] J. FERZIGER AND H. KAPER, *Mathematical Theory of Transport Processes in Gases*, North-Holland, Amsterdam, 1972.

[10] J. POLEWCZAK, *Global existence in L^1 for the modified nonlinear Enskog equation in R^3*, Tech. Report, Virginia Polytechnic Institute and State University, Blacksburg VA, 1988.

[11] G. TOSCANI, *On the Cauchy problem for the discrete Boltzmann equation with initial values in $L_1^+(R)$*, Tech. Report, Universita di Ferrara, 1987.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 647-659, May 1990
© 1990 Society for Industrial and Applied Mathematics
005

EXISTENCE OF STEADY-STATE SOLUTIONS FOR A ONE-PREDATOR–TWO-PREY SYSTEM*

NELA LAKOŠ†

Abstract. This paper discusses the existence of strictly positive solutions (in all three components) of the three-dimensional system of elliptic partial differential equations subject to Dirichlet boundary conditions, and models the situation in which a predator feeds on two-prey species. Results are obtained by the use of degree theory in cones, positive operators, and sub- and supersolution techniques.

Key words. predator-prey, competing species, system, bifurcation, degree, positive solutions, existence

AMS(MOS) subject classifications. 35J65, 92A17

0. Introduction. In this paper we study the three-dimensional system

$$\begin{aligned} -\Delta u &= u(a-u-cv-dz), \\ -\Delta v &= v(e+fu-v+gz), \\ -\Delta z &= z(\alpha-\beta u-\gamma v-z) \quad \text{in } \Omega, \\ u &= v = z = 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{0.1}$$

whose solutions are in fact steady-state (time-independent) solutions of a parabolic system. Here, Ω is a domain in $\mathfrak{R}^N$, with a smooth boundary. u, v, and z represent the population densities of certain species that co-inhabit the region Ω. u and z compete for the same food and are prey for the predator species v. All the parameters in (0.1) are assumed to be positive constants, except for e, which is also allowed to be negative. In the absence of one of the species, (0.1) reduces to a two-dimensional subsystem, either of competing species type ($v \equiv 0$) or of predator-prey type ($u \equiv 0$ or $z \equiv 0$). Solutions of either two-dimensional system are called extinction-state solutions of (0.1). Our goal is to establish sufficient conditions for the existence of componentwise strictly positive solutions of (0.1) (which we simply call positive solutions).

In § 2, we review some results obtained by various authors, of whom we mention a few—Leung [L1], Blat and Brown [BB], Dancer [D2], [D3]—who consider the existence of positive solutions of two-dimensional systems of both types. We also review (in modified form) results obtained by Cosner and Lazer [CL], Cantrell and Cosner [CC], McKenna and Walter [MW], and Leung [L2], regarding the uniqueness of positive solutions for competing species systems, and we obtain similar results for the predator-prey situation.

In § 3 we assume conditions under which the predator-prey subsystem ($z \equiv 0$) has a unique positive solution $(\bar{u}, \bar{v})$, and obtain a connected set (as α is varied) of positive solutions of (0.1) bifurcating from the branch $\{(\alpha, (\bar{u}, \bar{v}, 0)), \alpha \geqq 0\}$.

In § 4, we assume conditions under which the competing species subsystem has a unique positive solution $(\bar{u}, \bar{z})$ and, using e as a bifurcation parameter, we obtain a connected set of positive solutions of (0.1) bifurcating from the set $\{(e, (\bar{u}, 0, \bar{z})), e \geqq \nu\}$. As a consequence, we obtain our main result, Theorem 4.8.

Our main tool is degree theory with respect to cones and positive operator techniques, which were used by Dancer for the two-dimensional systems. We also use

* Received by the editors May 31, 1988; accepted for publication (in revised form) April 6, 1989.
† Department of Mathematics, Ohio State University, Columbus, Ohio 43210-1174.

sub- and supersolution techniques and variational characterization of eigenvalues, which are main tools in [CL].

The referee has informed us that some existence results for three species models were obtained by Korman and Leung [KL].

1. Preliminaries. Let λ_1 be the principle eigenvalue of

$$\begin{aligned} -\Delta\phi &= \lambda\phi \quad \text{in } \Omega, \\ \phi &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{1.1}$$

Let $m(x) \in C(\bar{\Omega})$ be such that $m(x_0) > 0$, for some $x_0 \in \Omega$. Then, by Theorem 1 of [HK], the boundary value problem

$$\begin{aligned} -\Delta\phi &= \lambda m\phi \quad \text{in } \Omega, \\ \phi &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{1.2}$$

has a principal eigenvalue, $\lambda_1(m) > 0$ and it is the only positive eigenvalue of (1.2) with a positive eigenfunction.

For every $p > 0$ such that $m + p > 0$ in Ω, define an operator $L_p := (-\Delta + p)^{-1}(m + p)$. Here $m + p$ denotes the Nemytskii operator associated with a function $m + p$; $(-\Delta + p)^{-1}$ denotes an inverse under Dirichlet boundary conditions; and $(-\Delta + p)^{-1}: C_0(\bar{\Omega}) \to C_0(\bar{\Omega})$ is a compact positive linear operator. Hence, L_p is the positive compact operator on $C_0(\bar{\Omega})$ and is irreducible (see [Sc, p. 269]), and therefore the spectral radius $r(L_p)$ is positive. $r(L_p)$ is also the only eigenvalue of L_p with a positive eigenfunction, by Theorem 3.2 of [Sc] (Krein-Rutman).

LEMMA 1.1. *If* $\lambda_1(m) = 1$, *then* $r(L_p) = 1$.

Proof. If $\lambda_1(m) = 1$, then there exists $\phi > 0$ such that (1.2) holds for $\lambda = 1$. This implies that $\phi = L_p\phi$. Therefore, $r(L_p) = 1$. □

Let $m, n \in C(\bar{\Omega})$ be such that $m < n$. Let $L_m = (-\Delta + p)^{-1}(m + p)$ and $L_n = (-\Delta + p)^{-1}(n + p)$.

LEMMA 1.2. $r(L_m) < r(L_n)$.

Proof. It is easy to show that $r(L_m) \leqq r(L_n)$. Assume that $r(L_m) = r(L_n) = r$. By the Krein-Rutman Theorem there exist $\varphi > 0$ and $\psi > 0$ such that $L_m\varphi = r\varphi$ and $L_n\psi = r\psi$ in Ω, and $\varphi = \psi = 0$ on $\partial\Omega$. This implies that

$$-\Delta r\varphi + pr\varphi = (m + p)\varphi \quad \text{in } \Omega, \tag{1.3}$$

$$-\Delta r\psi + pr\psi = (n + p)\psi \quad \text{in } \Omega. \tag{1.4}$$

Multiplying (1.3) by ψ and (1.4) by φ, then integrating over Ω and subtracting them, we obtain

$$0 = \int_\Omega (m - n)\varphi\psi \, dx < 0,$$

a contradiction. □

Consider the boundary value problem

$$\begin{aligned} -\Delta\phi &= \phi(a - \phi) \quad \text{in } \Omega, \\ \phi &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{1.5}$$

LEMMA 1.3. (i) *If* $a \leqq \lambda_1$, *then* (1.5) *has no nontrivial positive solution.*

(ii) *If* $a > \lambda_1$, *then there exists a unique positive solution* ϕ_a *of* (1.5) *and* $0 < \phi_a < a$. *Also,* $a < b$ *implies that* $\phi_a < \phi_b$.

Proof. See [BB, pp. 22–23] for the proof.

For $a > \lambda_1$ and $0 < k(<1)$, consider

$$\begin{aligned} -\Delta\phi &= \phi(a \pm k\phi_a - \phi) && \text{in } \Omega, \\ \phi &= 0 && \text{on } \partial\Omega. \end{aligned} \tag{1.6}$$

LEMMA 1.4. *$(1 \pm k)\phi_a$ is a unique positive solution of* (1.6).

Proof. The existence part is obvious. Assume there exists $u > 0$ such that (1.6) holds and $u \neq (1+k)\phi_a$. Let $\phi = (u - (1+k)\phi_a)$. Then ϕ satisfies the following boundary value problem

$$\begin{aligned} -\Delta\phi &= \phi(a - u - \phi_a) && \text{in } \Omega, \\ \phi &= 0 && \text{on } \partial\Omega. \end{aligned} \tag{1.7}$$

Then

$$\int_\Omega [-\Delta\phi - \phi(a - u - \phi_a)]\phi \, dx = 0. \tag{1.8}$$

On the other hand, since $\phi_a > 0$ is a solution of (1.5), it follows that for the eigenvalue problem

$$\begin{aligned} -\Delta\psi - \psi(a - \phi_a) &= \lambda\psi && \text{in } \Omega, \\ \psi &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{1.9}$$

zero is the lowest eigenvalue. It follows, due to variational characterization of the lowest eigenvalue (see [CH]), that

$$\int_\Omega [-\Delta\psi - \psi(a - \phi_a)]\psi \, dx \geqq 0 \quad \forall \psi \in C_0^2(\bar{\Omega}). \tag{1.10}$$

Formula (1.10), in particular, holds for ϕ. Now, (1.8) implies that $\int_\Omega u\phi^2 \, dx \leqq 0$, which is a contradiction. □

2. Two-species systems. In this section, we review some results for two-species systems. First, we consider the system modeling the one-predator–one-prey situation:

$$\begin{aligned} -\Delta u &= u(a - u - cv), \\ -\Delta v &= v(e + fu - v) && \text{in } \Omega, \\ u &= v = 0 && \text{on } \partial\Omega. \end{aligned} \tag{2.1}$$

For the proof of the following theorem, see [BB] or [D2].

THEOREM 2.1. *Assume that $a > \lambda_1$ and $e > \lambda_1$. For* (2.1) *there always exist two solutions, $(\phi_a, 0)$ and $(0, \phi_e)$. Let a^* be such that $\lambda_1(a^* - c\phi_e) = 1$. If $0 < c < 1$, then there exists a positive solution, (u, v), such that $u > 0$ and $v > 0$, for all $a \in (a^*, \infty)$, $(a^* \in (\lambda_1, e))$.*

Let

$$K = \frac{c^2 + f^2(1+f)^2}{(4+2cf)(1 - c(1+f))},$$

$$L = \frac{c^2 + f^2(1+f)^2}{4+2cf} + c(1+f).$$

It is easy to see that $0 < K < 1$ if and only if $0 < L < 1$.

Assume that $0 < K < 1$, $e > \lambda_1$ and define $\underline{a} := \inf\{a \in [\lambda_1, e];\ \phi_a \geqq L\phi_e\}$, and $\bar{a} := \sup\{a \in [e, \infty];\ K\phi_a \leqq \phi_e\}$. If we also assume that $a \in [\underline{a}, \bar{a}]$, the following lemma holds.

LEMMA 2.2. *Let* (u, v) *be a solution of* (2.1) *such that* $u>0$ *and* $v>0$. *Then* $u<\phi_a$ *and* $\phi_e<v<(1+f)\phi_e$. *If* $a\geqq e$, *then* $u\geqq(1-c(1+f))\phi_a$ *and if* $a\leqq e$, *then* $u\geqq(1-(c(1+f)/L))\phi_a$.

Proof. It has been proved in [D2] that $v\geqq\phi_e$ and $u\leqq\phi_a$. Since $v=0$ on $\partial\Omega$ and in Ω we have that $-\Delta v=v(e+fu-v)\leqq v(e+f\phi_a-v)$, it follows that v is a subsolution for (1.6). Since any big enough constant is a supersolution, Theorem 4.1 of [S] and Lemma 1.4 imply that $v\leqq(1+f)\phi_e$.

Assume that $a\geqq e$. Since $u=0$ on $\partial\Omega$ and in Ω we have that $-\Delta u=u(a-u-cv)\geqq u(a-u-c(1+f)\phi_a)$, it follows that u is a supersolution of (1.6), with $k=c(1+f)$. On the other hand, by the maximum principle (see [GT]) there exists an $\varepsilon>0$ such that $\varepsilon\phi_a<u$ and $\varepsilon<(1-c(1+f))$. Since we have that $-\Delta(\varepsilon\phi_a)=\varepsilon\phi_a(a-\phi_a)\leqq\varepsilon\phi_a(a-(\varepsilon+c(1+f)))\phi_a=\varepsilon\phi_a(a-c(1+f)\phi_a-\varepsilon\phi_a)$ in Ω and $\varepsilon\phi_a=0$ on $\partial\Omega$, $\varepsilon\phi_a$ is a subsolution of (1.6). Theorem 4.1 of [S] and Lemma 1.4 imply that $u\geqq(1-c(1+f))\phi_a$. We argue similarly for the case $a\leqq e$, the only difference being that $-\Delta u=u(a-u-cv)>u(a-u-c(1+f)\phi_e)\geqq u(a-u-(c(1+f)/L)\phi_a)$, in Ω. □

THEOREM 2.3. *If* $a\in[\underline{a}, \bar{a}]$, *there exists a unique positive solution* $(\bar{u}, \bar{v})$ *of* (2.1) *such that* $\bar{u}>0$ *and* $\bar{v}>0$.

Proof. Assume that there exist two different solutions (u_1, v_1) and (u_2, v_2) of (2.1) such that $u_i, v_i>0$, $i=1, 2$, for some $a\in[\underline{a}, \bar{a}]$. Let $p=u_1-u_2$, $q=v_1-v_2$. It is easy to check that (2.1) implies the following:

$$-\Delta p=p(a-u_1-cv_1)-u_2p-cu_2q,$$

$$-\Delta q=q(e+fu_1-v_1)+fv_2p-v_2q \quad \text{in } \Omega,$$

$$p=q=0, \quad \text{on } \partial\Omega.$$

Therefore,

$$\int_\Omega[-\Delta p-p(a-u_1-cv_1)]p\,dx+\int_\Omega(u_2p+cu_2q)p\,dx=0, \tag{2.2}$$

$$\int_\Omega[-\Delta q-q(e+fu_1-v_1)]q\,dx+\int_\Omega(v_2q-fv_2p)q\,dx=0. \tag{2.3}$$

Since (u_1, v_1) is a solution of (2.1), zero is the lowest eigenvalue for the following two eigenproblems:

$$-\Delta\phi-\phi(a-u_1-cv_1)=\alpha\phi \quad \text{in } \Omega,$$

$$\phi=0 \quad \text{on } \partial\Omega,$$

$$-\Delta\psi-\psi(e+fu_1-v_1)=\beta\psi \quad \text{in } \Omega,$$

$$\psi=0 \quad \text{on } \partial\Omega.$$

Arguing as in the proof of Lemma 1.4, we get that the first terms in both (2.2) and (2.3) are nonnegative. Therefore,

$$\int_\Omega(u_2p^2+(cu_2-fv_2)pq+v_2q^2)\,dx\leqq 0. \tag{2.4}$$

We would like to show that $D:=(cu_2-fv_2)^2-4u_2v_2<0$. That would prove that the

form in (2.4) is positive definite and imply that $p \equiv q \equiv 0$. In the case $a \geqq e$,

$$D = c^2u_2^2 + f^2v_2 - (4+2cf)u_2v_2 < c^2\phi_a^2 + f^2(1+f)^2\phi_a^2 - (4+2cf)(1-c(1+f))\phi_a\phi_e$$
$$\leqq [c^2 + f^2(1+f)^2 - (4+2cf)(1-c(1+f))K]\phi_a^2 = 0.$$

In the case where $a \leqq e$,

$$D < c^2\phi_e^2 + f^2(1+f)^2\phi_e^2 - (4+2cf)(1-(c(1+f)/L))\phi_a\phi_e$$
$$\leqq (c^2 + f^2(1+f)^2 - (4+2cf)(L-c(1+f))]\phi_e^2 = 0. \qquad \square$$

Next, we consider the two competing species system:

$$\begin{aligned} -\Delta u &= u(a-u-dz), \\ -\Delta z &= z(\alpha - \beta u - z) \quad \text{in } \Omega, \\ u &= z = 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{2.5}$$

It is clear that the first part of Theorem 2.1 holds for (2.5), as well. Assume that $a > \lambda_1$, $d < 1$, $\beta < 1$, and let

$$R = \frac{d^2+\beta^2}{(4-2\,d\beta)(1-d)} + \beta, \qquad Q = \frac{d^2+\beta^2}{(4-2\,d\beta)(1-\beta)} + d.$$

Assume that $0 < R < 1$ (this, obviously, implies that $0 < Q < 1$), and define $\alpha_1 := \inf\{\alpha \geqq \lambda_1,\ \phi_\alpha \geqq R\phi_a\}$, $\alpha_2 := \sup\{\alpha \geqq \lambda_1,\ Q\phi_\alpha \leqq \phi_a\}$. Assume that $d(1+\beta)/(1+d) < (a-\lambda_1)/a$. For the proof of the following theorem, see [L].

THEOREM 2.4. *Assume that* $\alpha \in (\lambda_1 + \beta a,\ (a-\lambda_1)/d)$. *Then there exists a positive solution* (u, z), *of* (2.5), *such that* $u > 0$ *and* $z > 0$.

LEMMA 2.5. *Assume that* $\alpha \in [\alpha_1, \alpha_2]$. *Then for any solution* (u, z) *of* (2.5), *such that* $u > 0$, $z > 0$, *the following holds. If* $\alpha \leqq a$, *then* $(1-d)\phi_a \leqq u \leqq \phi_a$, $(1-(\beta/R))\phi_\alpha \leqq z \leqq \phi_\alpha$; *if* $\alpha \geqq a$, *then* $(1-(d/Q))\phi_a \leqq u \leqq \phi_a$, $(1-\beta)\phi_\alpha \leqq z \leqq \phi_\alpha$.

Proof. Since $-\Delta u = u(a-u-dz) \geqq u(a-u-d\phi_\alpha)$, in the case $\alpha \geqq a$ we get that $-\Delta u \geqq u(a-u-(d/Q)\phi_a)$. In the case where $\alpha \leqq a$ we get that $-\Delta u \geqq u(a-u-d\phi_a)$. Since $-\Delta z = z(\alpha - \beta u - z) \geqq z(\alpha - \beta\phi_\alpha - z)$, in the case $\alpha \geqq a$ we get that $-\Delta z \geqq z(\alpha - \beta\phi_\alpha - z)$; in the case where $\alpha \leqq a$ we get that $-\Delta z \geqq z(\alpha - (\beta/R)\phi_\alpha - z)$. The rest follows, as in the proof of Lemma 2.2. $\square$

THEOREM 2.6. *If* $\alpha \in [\alpha_1, \alpha_2] \cap (\lambda_1 + \beta a,\ (a-\lambda_1)/d)$, *there exists a unique positive solution* $(\bar{u}, \bar{z})$ *of* (2.5) *such that* $\bar{u} > 0$ *and* $\bar{z} > 0$.

Proof. Assume that there exist two different solutions (u_1, z_1) and (u_2, z_2) such that $u_i > 0$, $z_i > 0$, $i = 1, 2$. Let $p = u_1 - u_2$, $q = z_1 - z_2$. Then

$$\begin{aligned} -\Delta p &= p(a-u_1-dz_1) - u_2p - du_2q, \\ -\Delta q &= q(\alpha - \beta u_1 - z_1) - \beta z_2 p - z_2 q \quad \text{in } \Omega, \\ p &= q = 0 \quad \text{on } \partial\Omega. \end{aligned}$$

As in the proof of Theorem 2.3, this implies that

$$\int_\Omega (u_2p^2 + (du_2 + \beta z_2)pq + z_2q^2)\,dx \leqq 0. \tag{2.6}$$

Let $\bar{D} = (du_2 + \beta z_2)^2 - 4u_2z_2$. Lemma 2.5 implies that $\bar{D} < 0$, which then implies that the form in (2.6) is positive definite. Therefore (2.6) can hold only for $p \equiv q \equiv 0$. $\square$

3. Three-species system with α as bifurcation parameter. In this section, we study the problem (0.1) and assume that all the parameters are fixed positive constants, except for α, which will serve as bifurcation parameter. We will deal with two different sets of assumptions:

$$\lambda_1 < e \leqq a \leqq \bar{a},$$

$$\beta + \gamma(1+f+g) \leqq K, \tag{3.1}$$

$$d + c(1+f+g) < 1,$$

$$\lambda_1 < \underline{a} \leqq a \leqq e,$$

$$\beta + \gamma(1+f+g) \leqq L, \tag{3.2}$$

$$d + c(1+f+g) \leqq L.$$

$\underline{a}$, $\bar{a}$, L, K were defined in the previous section. All the results will be stated for (3.1), and statements in parentheses will refer to (3.2). In the absence of parentheses, the result holds for both cases.

LEMMA 3.1. *Let* (u, v, z) *be a solution of* (0.1), *such that* $u, v, z \geqq 0$. *Then*

(i) $u \leqq \phi_a$;

(ii) $v \leqq (1+f+g)\phi_\eta$, $\eta = \max\{\alpha, a, e\}$; *if* $v \not\equiv 0$, $\phi_e \leqq v$;

(iii) $z \leqq \phi_\alpha$ *and if* $z \not\equiv 0$ *and* $\alpha \geqq a, e$, *then* $(1-[\beta+\gamma(1+f+g)])\phi_\alpha \leqq z$.

Proof. We proceed as in the proof of Lemma 2.5. The crucial inequalities are:

(ii) $$-\Delta v = v(e+fu+gz-v) \leqq v(e+f\phi_a+g\phi_\alpha-v) \leqq v(\eta+(f+g)\phi_\eta-v);$$

(iii) $$-\Delta z = z(\alpha-\beta u-\gamma v-z) \geqq z(\alpha-\beta\phi_a-\gamma(1+f+g)\phi_\alpha-z)$$

$$= z(\alpha-[\beta+\gamma(1+f+g)]\phi_\alpha-z).$$ □

We now establish an appropriate setting that will enable us to transform problem (0.1) into a fixed-point equation. To this end, let $E = [C_0(\bar{\Omega})]^3$, be a Banach space with the norm $\|u\| = \|(u_1, u_2, u_3)\| = \max\{\|u_i\|_0,\ i = 1, 2, 3\}$. Let $P = [C_0(\bar{\Omega})^+]^3$ be a cone of positive functions in E. For all $\alpha > 0$, define a set $T_\alpha \subset P$ by $T_\alpha := \{(u, v, z) \in P,\ u \leqq 2a,\ v \leqq 2\eta(1+f+g),\ z \leqq 2\alpha\}$. By Lemma 3.1, all positive solutions of (0.1) lie in the interior (with respect to relative topology on P) of T_α. Also, there exists a continuous, nondecreasing function of α, $p(\alpha)$, such that $a-u-cv-dz+p(\alpha) > 0$, $e+fu-v+gz+p(\alpha) > 0$, $\alpha-\beta u-\gamma v-z+p(\alpha) > 0$, for all $(u, v, z) \in T_\alpha$. This enables us to define $A(\alpha, \cdot)$, an operator on T_α, by

$$A(\alpha, (u, v, z)) := (-\Delta+p(\alpha))^{-1}(u(a-u-cv-dz+p(\alpha)), v(e+fu-v+gz+p(\alpha)),$$

$$z(\alpha-\beta u-\gamma v-z+p(\alpha))). \tag{3.3}$$

$A(\alpha, \cdot)$: $T_\alpha \to P$; it is completely continuous and Fréchet differentiable, and fixed points of $A(\alpha, \cdot)$ are solutions of (0.1). Since we are interested in positive solutions of (0.1), we will study equations of the form

$$A(\alpha, (u, v, z)) = (u, v, z) \tag{3.4}$$

instead of (0.1), i.e., we will study fixed points of a one-parameter family of completely continuous maps. This family, A: $T \to P$, where $T = \bigcup_{\alpha \geqq 0}\{\alpha\} \times T_\alpha$, $T \subset \mathfrak{R}^+ \times P$, is a

completely continuous map. Therefore, the solution set S of (3.4) defined by

$$S := \{(\alpha, (u, v, z)) \in T, A(\alpha, (u, v, z)) = (u, v, z)\}$$

is locally compact.

Let $\delta = 1/d(1-[\beta+\gamma(1+f+g)])$ and let α^* be such that $\lambda_1(\alpha^* - \gamma\phi_e) = 1$.

LEMMA 3.2. *If $\alpha \geqq a\delta (\alpha \geqq e\delta)$ or $\alpha \leqq \alpha^*$, then $A(\alpha, \cdot)$ has no fixed point with all three components nontrivial.*

Proof. Assume that $\alpha \leqq \alpha^*$ and that there exists $(u, v, z) \in T_\alpha$, a fixed point of $A(\alpha, \cdot)$ such that $u > 0$, $v > 0$, and $z > 0$. Let $p = p(\alpha)$. Since $z = (-\Delta+p)^{-1}(\alpha - \beta u + \gamma v - z + p)z$, it follows that $r((-\Delta+p)^{-1}(\alpha - \beta u - \gamma v - z + p)) = 1$. On the other hand,

$$\alpha - \beta u - \gamma v - z < \alpha - \gamma v < \alpha - \gamma\phi_e \leqq \alpha^* - \gamma\phi_e. \tag{3.5}$$

Also, Lemma 1.1 implies that $r((-\Delta+p)^{-1}(\alpha^* - \gamma\phi_e + p)) = 1$. Now, (3.5) and Lemma 1.2 imply that $r((-\Delta+p)^{-1}(\alpha - \beta u - \gamma v - z + p)) < 1$, which is a contradiction.

Let $\alpha \geqq a\delta$ and let $(u, v, z) \in T_\alpha$, $u > 0$, $v > 0$, $z > 0$, be a fixed point of $A(\alpha, \cdot)$. Since $a - u - cv - dz + p < a - u - d(1-\beta-\gamma(1+f+g))\phi_\alpha + p = a - u - 1/\delta\phi_\alpha + p$, and $u = (-\Delta+p)^{-1}(a - u - cv - dz + p)u$, it follows that

$$u \leqq (-\Delta+p)^{-1}\left(a - u - \frac{1}{\delta}\phi_\alpha + p\right)u. \tag{3.6}$$

This implies that $r((-\Delta+p)^{-1}(a - u - 1/\delta\phi_\alpha + p)) \geqq 1$, by definition of spectral radius. On the other hand,

$$\begin{aligned} a - u - \frac{1}{\delta}\phi_\alpha < a - \frac{1}{\delta}\phi_\alpha &= \frac{1}{\delta}(a\delta - \phi_\alpha) \\ &< \alpha - \phi_\alpha. \end{aligned}$$

Arguing as in the previous case, we get

$$r\left((-\Delta+p)^{-1}\left(a - u - \frac{1}{\delta}\phi_\alpha + p\right)\right) < r((-\Delta+p)^{-1}(\alpha - \phi_\alpha + p)) = 1,$$

a contradiction.

In the case $\lambda_1 < a \leqq e$ and $\alpha \geqq e\delta$, the crucial inequalities are

$$\begin{aligned} u &= (-\Delta+p)^{-1}(a - u - cv - dz + p)u \\ &\leqq (-\Delta+p)^{-1}\left(e - u - \frac{1}{\delta}\phi_\alpha + p\right)u, \end{aligned}$$

$$e - u - \frac{1}{\delta}\phi_\alpha < \frac{1}{\delta}(e\delta - \phi_\alpha) < \alpha - \phi_\alpha. \qquad \square$$

Let $(\bar{u}, \bar{v})$ be a unique positive solution of (2.1) such that $\bar{u} > 0$ and $\bar{v} > 0$, guaranteed by Theorem 2.3.

LEMMA 3.3. *There exists a unique $\bar{\alpha} > 0$ such that*

$$r((-\Delta + p(\bar{\alpha}))^{-1}(\bar{\alpha} - \beta\bar{u} - \gamma\bar{v} + p(\bar{\alpha}))) = 1, \tag{3.7}$$

$\bar{\alpha} \in (\alpha^*, e) (\bar{\alpha} \in (\alpha^*, a))$. Also, $r((-\Delta + p(\alpha))^{-1}(\alpha - \beta\bar{u} - \gamma\bar{v} + p(\alpha)))$ is less than 1, if $\alpha < \bar{\alpha}$, and greater than 1, if $\bar{\alpha} < \alpha$.

Proof. If $\alpha \leqq \alpha^*$, then $\alpha - \beta\bar{u} - \gamma\bar{v} < \alpha - \gamma\phi_e \leqq \alpha^* - \gamma\phi_e$. Arguing as in Lemma 3.2, we get that

$$r((-\Delta + p(\alpha))^{-1}(\alpha - \beta\bar{u} - \gamma\bar{v} + p(\alpha))) < r((-\Delta + p(\alpha))^{-1}(\alpha^* - \gamma\phi_e + p(\alpha))) = 1.$$

If $\alpha \geqq e$, then

$$\begin{aligned}\alpha - \beta\bar{u} - \gamma\bar{v} &> \alpha - \beta\phi_a - \gamma(1+f)\phi_e > e - (\beta + \gamma(1+f))\phi_a \\ &\geqq e - \frac{\beta + \gamma(1+f)}{K}\phi_e \\ &\geqq e - \phi_e.\end{aligned}$$

This implies that $r((-\Delta + p(\alpha))^{-1}(\alpha - \beta\bar{u} - \gamma\bar{v} + p(\alpha))) > 1$. (In the case where $\lambda_1 < a \leqq e$ and $\alpha \geqq a$, the crucial inequalities are

$$\begin{aligned}\alpha - \beta\bar{u} - \gamma\bar{v} &> \alpha - \beta\phi_a - \gamma(1+f)\phi_e > \alpha - (\beta + \gamma(1+f))\phi_e \\ &> a - \frac{\beta + \gamma(1+f)}{L}\phi_a > a - \phi_a.)\end{aligned}$$

Let $r(\alpha)$: $\mathfrak{R}^+ \to \mathfrak{R}^+$ be defined by

$$r(\alpha) := r((-\Delta + p(\alpha))^{-1}(\alpha - \beta\bar{u} - \gamma\bar{v} + p(\alpha))).$$

Then Theorem 4.3.1 and § 4.3.5 of [K] imply that $r(\alpha)$ is a continuous function. Since $r(\alpha^*) < 1$ and $r(e) > 1$ ($r(a) > 1$), it follows that there exists an $\bar{\alpha} \in (\alpha^*, e)$, ($\bar{\alpha} \in (\alpha^*, a)$), such that $r(\bar{\alpha}) = 1$, i.e., (3.7) holds. Assume there exists α', $\alpha' \neq \bar{\alpha}$ such that $r(\alpha') = 1$. Then there exist $\phi > 0$ and $\psi > 0$ such that (1.2) holds, respectively, for $m = \bar{\alpha} - \beta\bar{u} - \gamma\bar{v}$ and $n = \alpha' - \beta\bar{u} - \gamma\bar{v}$. It follows that $\int_\Omega -\Delta\phi \cdot \psi \, dx = \int_\Omega m\psi\phi \, dx$ and $\int_\Omega -\Delta\psi \cdot \phi \, dx = \int_\Omega n\psi\phi \, dx$. Subtracting these two equalities, we get that $0 = \int_\Omega (m-n)\phi\psi \, dx \neq 0$, which is a contradiction.

Since $\bar{\alpha}$ is the only value of α such that $r(\alpha) = 1$, and since $r(\alpha)$ is a continuous function and $r(\alpha^*) < 1$, $r(e) > 1$ ($r(a) > 1$), it follows that $r(\alpha) < 1$ if $\alpha < \bar{\alpha}$, and that $r(\alpha) > 1$ if $\alpha > \bar{\alpha}$. □

Remark 3.4. Let L_α be the linearization of $A(\alpha, \cdot)$ at the point $(\bar{u}, \bar{v}, 0) \in T_\alpha$. ($(\bar{u}, \bar{v}, 0)$ is, obviously, a fixed point of $A(\alpha, \cdot)$.) Then

$$\begin{aligned}L_\alpha(l, k, h) = (-\Delta + p(\alpha))^{-1}((a - 2\bar{u} - c\bar{v} + p(\alpha))l - c\bar{u}k - d\bar{u}h, f\bar{v}l \\ + (e + f\bar{u} - 2\bar{v} + p(\alpha))k + g\bar{v}h, (\alpha - \beta\bar{u} - \gamma\bar{v} + p(\alpha))h).\end{aligned}$$

Next, we would like to compute the fixed-point index of $A(\alpha, \cdot)$ at the point $(\bar{u}, \bar{v}, 0)$ relative to the cone P. Let $i(A(\alpha, \cdot), y)$ denote the fixed-point index of $A(\alpha, \cdot)$ at y with respect to the cone P.

LEMMA 3.5. *$i(A(\alpha, \cdot), (\bar{u}, \bar{v}, 0))$ is equal to zero if $\alpha > \bar{\alpha}$, and equal to ± 1 if $\alpha < \bar{\alpha}$.*

Proof. We will use the notation from [D1]. Let $y = (\bar{u}, \bar{v}, 0)$. Then $W_y = C_0(\bar{\Omega}) \times C_0(\bar{\Omega}) \times C_0(\bar{\Omega})^+$. First we have to show that L_α has no eigenvector in W_y corresponding to eigenvalue 1. Assume, on the contrary, that there exists $(l, k, h) \in W_y$ such that

$$L_\alpha(l, k, h) = (l, k, h). \tag{3.8}$$

Assume that $h \not\equiv 0$. Equation (3.8) implies that $(-\Delta + p(\alpha))^{-1}(\alpha - \beta\bar{u} - \gamma\bar{v} + p(\alpha))h = h$, and therefore, $r((-\Delta + p(\alpha))^{-1}(\alpha - \beta\bar{u} - \gamma\bar{v} + p(\alpha))) = 1$. This contradicts the assumption that $\alpha \neq \bar{\alpha}$. Hence $h \equiv 0$. If $l \not\equiv 0 \not\equiv k$, then (3.8) implies that $-\Delta l - (a - 2\bar{u} - c\bar{v})l + c\bar{u}k = 0$, and $-\Delta k - f\bar{v}l - (e + f\bar{u} - 2\bar{v})k = 0$, in Ω, and $l = k = 0$, on $\partial\Omega$. Multiplying the

first equality by l, and the second by k and then integrating over Ω, we get

$$\int_\Omega [-\Delta l \cdot l - (a - \bar{u} - c\bar{v})l^2]\, dx + \int_\Omega [\bar{u}l^2 + c\bar{u}kl]\, dx = 0,$$

$$\int_\Omega [-\Delta k \cdot k - (e + f\bar{u} - \bar{v})k^2]\, dx + \int_\Omega [\bar{v}k^2 - f\bar{v}lk]\, dx = 0.$$

Arguing as in the proof of Lemma 2.5, we get that

$$\int_\Omega [\bar{u}l^2 + \bar{v}k^2 + (c\bar{u} - f\bar{v})lk]\, dx \leqq 0.$$

The form above is equal to the form in (2.4), which is positive definite. Hence $l \equiv k \equiv 0$.

Following the notation of [D1], we find that $S_y = C_0(\bar{\Omega}) \times C_0(\bar{\Omega}) \times \{0\}$. Let $M = \{0\} \times \{0\} \times C_0(\bar{\Omega})$. Then M is a closed complement of S_y in E. Let Π be the projection onto M, and let M_α denote the restriction of L_α to M. Then $\Pi \circ M_\alpha = (-\Delta + p(\alpha))^{-1}(\alpha - \beta\bar{u} - \gamma\bar{v} + p(\alpha))$. Now, Theorem 1 (with Remark 2) and Lemma 2 (with Remark 3) of [D1] imply that $i(A(\alpha, \cdot), (\bar{u}, \bar{v}, 0)) = i(L_\alpha, 0, S_y) = \pm 1$, if $\alpha < \bar{\alpha}$, and $i(A(\alpha, \cdot), (\bar{u}, \bar{v}, 0)) = 0$, if $\alpha > \bar{\alpha}$. □

Let $C_0 = \{(\alpha, (\bar{u}, \bar{v}, 0)), \alpha \geqq 0\}$. Then $C_0 \subset S$; it is a continuum (a closed, connected set) in S, and therefore in $\Re^+ \times P$.

LEMMA 3.6. *$\bar{\alpha}$ is a bifurcation point for* (3.4) *with respect to C_0. It is the only one.*

Proof. Assume that $\bar{\alpha}$ is not a bifurcation point for C_0. Then there exists an interval $[B, C]$ such that $\bar{\alpha} \in [B, C]$ and an open set, U, $U \subset [B, C] \times P \cap T$ (in relative topology) such that $\bar{U} \cap S = [B, C] \times \{(\bar{u}, \bar{v}, 0)\}$ and $\partial U \cap S = \varnothing$. Let $U_\alpha = \{(u, v, z), (\alpha, (u, v, z)) \in U\}$. Then, by the general homotopy invariance property of the fixed-point index (see [A, Thm. 11.3]), $i(A(B, \cdot), U_B) = i(A(C, \cdot), U_C)$. On the other hand, by the excision property and definition of local index (see [A, p. 659]), $i(A(B, \cdot), U_B) = i(A(B, \cdot), (\bar{u}, \bar{v}, 0)) = \pm 1$, and $i(A(C, \cdot), U_C) = i(A(C, \cdot), (\bar{u}, \bar{v}, 0)) = 0$, since $B < \bar{\alpha}$, and $C > \bar{\alpha}$, which is a contradiction.

Assume that there exists another bifurcation point α, $\alpha \neq \bar{\alpha}$. Then there exists a sequence in $S \setminus C_0$, $\{(\alpha_n, (u_n, v_n, z_n)), n \in \aleph\}$ such that $\alpha_n \to \alpha$, $u_n \to \bar{u}$, $v_n \to \bar{v}$, $z_n \to 0$. Also, there exists $N \in \aleph$ such that $u_n > 0$, $v_n > 0$, for all $n \geqq N$.

Assume that $z_n = 0$, for some $n \geqq N$. Then (u_n, v_n) is a solution of (2.1) and therefore $u_n = \bar{u}$, $v_n = \bar{v}$, and hence $(\alpha_n, (u_n, v_n, z_n)) \in C_0$, contrary to the assumption. So, $z_n > 0$ for all $n \geqq N$. The following holds:

$$\frac{z_n}{\|z_n\|} = (-\Delta)^{-1}\left(\frac{z_n}{\|z_n\|}(\alpha_n - \beta u_n - \gamma v_n - z_n)\right), \qquad n \geqq N. \tag{3.9}$$

The sequence on the right-hand side converges for some subsequence (which we relabel as the original one), since $(-\Delta)^{-1}$ is a compact operator and the sequence in the parentheses is bounded. Therefore, the left-hand side of (3.9) converges also, to some z of norm 1. Passing to the limit in (3.9), we get that $z = (-\Delta)^{-1}(z(\alpha - \beta\bar{u} - \gamma\bar{v}))$, which implies that $z = (-\Delta + p(\alpha))^{-1}(\alpha - \beta\bar{u} - \gamma\bar{v} + p(\alpha))z$, and by Lemma 3.3, it follows that $\alpha = \bar{\alpha}$, contrary to the assumption. □

Let $\Sigma = (S \setminus C_0) \cup \{(\bar{\alpha}, (\bar{u}, \bar{v}, 0))\}$. Σ is a closed subset of S, by Lemma 3.6.

LEMMA 3.7. *Σ contains an unbounded continuum $\tilde{C}$, bifurcating from C_0 at $\bar{\alpha}$.*

Proof. Let $\tilde{C}$ be the component of Σ containing $(\bar{\alpha}, (\bar{u}, \bar{v}, 0))$. Assume that $\tilde{C}$ is bounded. Then there exists $\mu > \bar{\alpha}$ such that $\tilde{C} \subset [0, \mu] \times P \cap T$ and $\tilde{C} \cap \{\mu\} \times T_\mu = \varnothing$. Let $X = S \cap [0, \mu] \times P$. X is obviously a compact topological space. Let $Y = \tilde{C} \cup C_0 \cap X$, $Z = S \cap (\{\mu\} \times P \cup \{0\} \times P) \setminus Y$. Then Y and Z are nonempty, disjoint, closed subsets

of X (for example $(\mu, (0, \phi_e, 0)) \in Z$). By Whyburn's Lemma (see [W]) there exist two compact sets V and W, such that $Y \subset V$, $Z \subset W$, $V \cap W = \emptyset$, $V \cup W = X$. This implies that there exists an open set U in $[0, \mu] \times P$, such that $U \supset V \supset Y$, $U \cap W = \emptyset$ and $\partial U \cap S = \emptyset$. Therefore, $i(A(\alpha, \cdot), U_\alpha)$ is well defined for all $\alpha \in [0, \mu]$, and it is constant (with respect to α) by the homotopy invariance principle (see [A, Thm. 11.3]). On the other hand, $i(A(\mu, \cdot), U_\mu) = i(A(\mu, \cdot), (\bar{u}, \bar{v}, 0)) = 0$, since $\mu > \bar{\alpha}$. Therefore, $i(A(0, \cdot), U_0) = 0$. Since $i(A(0, \cdot), (\bar{u}, \bar{v}, 0)) = \pm 1$, it follows that $\tilde{C} \cap \{0\} \times P \neq \emptyset$. Let $(0, (u, v, z)) \in \tilde{C} \cap \{0\} \times P$. If $z \not\equiv 0$, then (3.2) implies that $-\Delta z = -z(\beta u + \gamma v + z)$, in Ω and $z = 0$, on $\partial\Omega$, which contradicts the maximum principle. Hence $z \equiv 0$. Therefore, (u, v) is a solution of (2.1), different from $(\bar{u}, \bar{v})$. So, at least one of the components must be zero. It cannot be v, since $\tilde{C}$ is a continuum and positive v-components are bounded away from zero. So, $u \equiv 0$, and $v = \phi_e$. Since $\tilde{C} \cap \{0\} \times P = \{(0, (0, \phi_e, 0))\}$, $\tilde{C}$ must contain the whole unbounded continuum $\{(\alpha, (0, \phi_e, 0)), \alpha \geqq 0\}$, contrary to the assumption. □

LEMMA 3.8. *Let C be the component of $\tilde{C} \cap [0, a] \times P$ ($\tilde{C} \cap [0, e] \times P$), containing $(\bar{\alpha}, (\bar{u}, \bar{v}, 0))$. Then*

(i) *$u > 0$, $v > 0$, $z > 0$, for all $(\alpha, (u, v, z)) \in C \setminus \{(\bar{\alpha}, (\bar{u}, \bar{v}, 0))\}$;*

(ii) *$C \cap [0, \alpha^*] \times P = \emptyset$;*

(iii) *$C \cap \{\alpha\} \times P \neq \emptyset$, for all $\alpha \in [\bar{\alpha}, a]$, ($[\bar{\alpha}, e]$).*

Proof. Part (i) must hold for some neighborhood of $(\bar{\alpha}, (\bar{u}, \bar{v}, 0))$ in C. Take any $(\alpha, (u, v, z)) \in S$ such that $\alpha \leqq a$ and $u > 0$. Since $-\Delta u = u(a - u - cv - dz) \geqq u(a - u - c(1+f+g)\phi_a - d\phi_\alpha) \geqq u(a - u - [c(1+f+g) + d]\phi_a)$, in Ω and $u = 0$ on $\partial\Omega$, it follows, as in the proof of Lemma 2.2, that

$$u \geqq (1 - [c(1+f+g) + d])\phi_a. \tag{3.10}$$

Assume that (i) does not hold. Then there exists $(\alpha, (u, v, z)) \in C$ such that either $u \equiv 0$, or $v \equiv 0$, or $z \equiv 0$. Assume that $u \equiv 0$. Then (3.10) implies that $u \equiv 0$, for all $(\alpha(u, v, z)) \in C$, since C is a continuum, which is a contradiction. Assume that $v \equiv 0$. Then $v \equiv 0$, for all $(\alpha, (u, v, z)) \in C$, since for positive v, $v > \phi_e$ and C is a continuum. This is, again, a contradiction. If $z \equiv 0$, then $u = \bar{u}$, $v = \bar{v}$ and, therefore $\alpha = \bar{\alpha}$, which proves (i). Part (ii) follows from Lemma 3.2 and (iii) follows from Lemmas 3.3 and 3.7. □

4. Three-species system with e as bifurcation parameter. In this section we assume that all the parameters in (0.1) are fixed positive constants, except for e. Since we follow the approach of the previous section, we try to avoid repetition by omitting or sketching the proofs only. We assume the following:

$$\begin{aligned} &a > \lambda_1, \\ &d, \beta, d\frac{1+\beta}{1+d} < \frac{a-\lambda_1}{a}, \\ &0 < R < 1, \\ &\alpha \in [\alpha_1, \alpha_2] \cap \left(\lambda_1 + \beta a, \frac{a-\lambda_1}{d}\right), \\ &\beta + \gamma(1+f+g) < R, \qquad d + c(1+f+g) < Q. \end{aligned} \tag{4.1}$$

All the constants in (4.1) have been defined in § 2. Let E and P be as in the previous section. Let $\nu = \lambda_1 - 2\alpha/R$. Since Lemma 3.1 holds, the set $T_e \subset P$, defined below, for all $e \in [\nu, \infty)$, contains in its interior all the positive solutions of (0.1):

$$T_e := \{(u, v, z) \in P, u \leqq 2a, v \leqq 2\eta(1+f+g), z \leqq 2\alpha\}, \qquad \eta = \max\{\alpha, a, e\}.$$

Let $p(e)$ be a continuous, nondecreasing function on $[\nu, \infty)$, such that $a-u-cv-dz+p(e)>0$, $e+fu-v+gz+p(e)>0$, $\alpha-\beta u-\gamma v-z+p(e)>0$, for all $(u, v, z)\in T_e$. Define an operator $A(e, \cdot)$ on T_e by (3.3), by writing $p(e)$ instead of $p(\alpha)$. We consider a one-parameter family of fixed-point equations on $T:=\bigcup_{e\geqq\nu}\{e\}\times T_e$:

$$A(e, (u, v, z))=(u, v, z). \tag{4.2}$$

A solution set S of (4.2), defined by $S:=\{(e, (u, v, z))\in T, A(e, (u, v, z))=(u, v, z)\}$, is locally compact.

LEMMA 4.1. *Let (u, v, z) be a solution of* (4.2). *Then $e\geqq\alpha/\gamma$ implies that $z\equiv 0$, and $e\geqq a/c$ implies that $u\equiv 0$.*

Proof. Let (u, v, z) be a solution of (4.2) such that $e\geqq\alpha/\gamma$ and $z>0$. Then $r((-\Delta+p(e))^{-1}(\alpha-\beta u-\gamma v-z+p(e)))=1$. On the other hand, $\alpha-\beta u-\gamma v-z<\alpha-\gamma\phi_e=\gamma(\alpha/\gamma-\phi_e)\leqq\gamma(e-\phi_e)\leqq e-\phi_e$, which implies a contradiction.

If (u, v, z) is a solution of (4.2) such that $e\geqq a/c$ and $u>0$, then $r((-\Delta+p(e))^{-1}(a-u-cv-dz+p(e)))=1$. On the other hand, $a-u-cv-dz<a-c\phi_e=c((a/c)-\phi_e)\leqq c(e-\phi_e)\leqq e-\phi_e$, which implies a contradiction. □

LEMMA 4.2. *Assume that $e\leqq\min\{a, \alpha\}$ and let (u, v, z) be a solution of* (4.2) *such that $u>0$ and $z>0$. Then*

$$z\geqq\left(1-\frac{\beta+\gamma(1+f+g)}{R}\right)\phi_\alpha \quad \textit{and} \quad u\geqq\left(1-\frac{d+c(1+f+g)}{Q}\right)\phi_a.$$

Proof. We argue as in the proof of Lemma 2.2. The main inequalities are

$$\begin{aligned}-\Delta z&=z(\alpha-\beta u-\gamma v-z)\geqq z(\alpha-(\beta+\gamma(1+f+g))\phi_\eta-z)\\&\geqq z(\alpha-((\beta+\gamma(1+f+g))/R)\phi_\alpha-z),\end{aligned}$$

and

$$\begin{aligned}-\Delta u&=u(a-u-cv-dz)>u(a-u-(d+c(1+f+g))\phi_\eta)\\&\geqq u(a-u-((d+c(1+f+g))/Q)\phi_a).\end{aligned}$$

□

Let $(\bar u, \bar z)$ be a unique positive solution of (2.5), guaranteed by Theorem 2.6.

LEMMA 4.3. *There exists a unique $\bar e\in[\nu, \infty)$ such that*

$$r(\bar e):=r((-\Delta+p(\bar e))^{-1}(\bar e+f\bar u+g\bar z+p(\bar e)))=1.$$

Also, $r(e)<1$ for $e<\bar e$, and $r(e)>1$ for $e>\bar e$.

Proof. Let $e=\lambda_1-(f/R+g)\alpha$. Then

$$e+f\bar u+g\bar z<e+f\phi_a+g\phi_\alpha<e+((f/R+g)\phi_\alpha<e+(f/R)+g)\alpha=\lambda_1.$$

Arguing as in Lemma 3.3, we get that $r((-\Delta+p(e))^{-1}(e+f\bar u+g\bar z+p(e)))<1$.

On the other hand, for $e=\lambda_1$ we get that $r((-\Delta+p(e))^{-1}(e+f\bar u+g\bar z+p(e)))>1$. The rest follows as in the proof of Lemma 3.3. □

As in the previous section, we would like to compute the fixed-point index of $A(e, \cdot)$ at the point $(\bar u, 0, \bar z)$, relative to the cone P, as the parameter e is varied. To prove the following, we proceed step by step as in the proof of Lemma 3.5.

LEMMA 4.4. *$i(A(e, \cdot), (\bar u, 0, \bar z))$ is equal to ± 1 if $e<\bar e$, and is equal to zero if $e>\bar e$.*

Let $C_0=\{(e, (\bar u, 0, \bar z)), e\geqq\nu\}$. $C_0\subset S$ is a continuum in S and in $[\nu, \infty)\times P$.

LEMMA 4.5. *$\bar e$ is a bifurcation point for* (4.2) *with respect to C_0. It is the only one.*

Proof. Follow the proof of Lemma 3.6. □

Let $\Sigma=(S\backslash C_0)\cup\{(\bar e, (\bar u, 0, \bar z))\}$. Lemma 4.5 implies that Σ is a closed subset of S.

LEMMA 4.6. *Σ contains an unbounded continuum $\tilde C$ bifurcating from C_0 at $\bar e$.*

Proof. We proceed as in the proof of Lemma 3.7. Let $\tilde C$ be the component of Σ containing $(\bar e, (\bar u, 0, \bar z))$. Assume that $\tilde C$ is bounded. Then there exists $\mu>e$ such that

$\tilde{C} \subset [\nu, \mu] \times P$ and $\tilde{C} \cap \{\mu\} \times P = \emptyset$. Let $X = S \cap [\nu, \mu] \times P$, $Y = \tilde{C} \cup C_0 \cap X$, $Z = S \cap (\{\mu\} \times P \cup \{\nu\} \times P) \setminus Y$. Arguing as in the proof of Lemma 3.7, we can find an open set U in $[\nu, \mu] \times P$ such that $Y \subset U$, $\bar{U} \cap Z = \emptyset$, and $\partial U \cap S = \emptyset$. So, $i(A(e, \cdot), U_e)$ is well defined for all $e \in [\nu, \mu]$ and constant with respect to e. It follows that $i(A(\mu, \cdot), U_\mu) = i(A(\mu, \cdot), (\bar{u}, 0, \bar{z})) = 0$, since $\mu > \bar{e}$. Therefore, $i(A(\nu, \cdot), U_\nu) = 0$. Since $i(A(\nu, \cdot), (\bar{u}, 0, \bar{z})) = \pm 1$, and $U_\nu \cap S = (\tilde{C} \cap \{\nu\} \times P) \cup \{(\nu, (\bar{u}, 0, \bar{z}))\}$, it follows that $\tilde{C} \cap \{\nu\} \times P \neq \emptyset$. Let $(\nu, (u, v, z)) \in \tilde{C} \cap \{\nu\} \times P$. If $v \not\equiv 0$, then $-\Delta v = v(\nu + fu - v + gz) \leqq -v^2$ in Ω, and $v \equiv 0$ on $\partial\Omega$. This contradicts the maximum principle (see [GT]). Since $v \equiv 0$, it follows that (u, z) is a solution of (2.5), and therefore $u \equiv 0$ or $z \equiv 0$. So, $\tilde{C} \cap \{\nu\} \times P$ contains either $(\nu, (0, 0, 0))$, $(\nu, (0, 0, \phi_\alpha))$ or $(\nu, (\phi_a, 0, 0))$. In either case, $\tilde{C}$ must contain a whole unbounded continuum, for example, $\{(e, (0, 0, 0)), e \geqq \nu\}$, contrary to the assumption. □

LEMMA 4.7. *Let C be the component of $\tilde{C} \cap [\nu, \min\{a, \alpha\}] \times P$, containing $\{(\bar{e}, (\bar{u}, 0, \bar{z}))\}$. Then*

(i) *$u > 0$, $v > 0$, $z > 0$ for all $(e, (u, v, z)) \in C \setminus \{(\bar{e}, (\bar{u}, 0, \bar{z}))\}$;*

(ii) *$C \cap \{e\} \times P \neq \emptyset$ for all $e \in (\bar{e}, \min\{a, \alpha\}]$.*

Proof. Part (i) must hold in some neighborhood of $(\bar{e}, (\bar{u}, 0, \bar{z}))$ in C. Lemma 4.2 implies that $u > 0$ and $z > 0$ for all $(e, (u, v, z)) \in C \setminus \{(\bar{e}, (\bar{u}, 0, \bar{z}))\}$. Then $v > 0$, since (u, z) solves (2.5). Part (ii) is obvious. □

Lemmas 4.7 and 3.8 imply the following theorem.

THEOREM 4.8. *Assume that* (3.1) *or* (3.2) *holds. Then there exists a positive solution of* (0.1) *for all $\alpha \in (\bar{\alpha}, \max\{a, e\})$. If* (4.1) *holds instead, the same is true for all $e \in (\bar{e}, \min\{a, \alpha\})$.*

By Lemma 3.3, $\bar{\alpha} \in (\alpha^*, \min\{a, e\})$ and an estimate for $\bar{e}$ is provided by the following remark.

Remark 4.9. Let ω be an eigenfunction of (1.1) associated with λ_1, such that $\|\omega\| = 1$, $\omega > 0$. Define $\kappa := \int_\Omega \omega^3\, dx / \int_\Omega \omega^2\, dx$. The following estimate holds:

$$\bar{e} < \lambda_1[1 - \kappa(a - \lambda_1)(f(Q - d) + g(R - \beta))].$$

Proof. Let $\xi = f(Q - d) + g(R - \beta)$. Then, there exists a unique e such that

$$r((-\Delta + p(e))^{-1}(e + \xi\phi_a + p(e))) = 1. \tag{4.3}$$

We can prove this by arguing as in the proof of Lemma 3.3. Since $e + f\bar{u} + g\bar{z} > e + f(1 - d/Q)\phi_a + g(1 - \beta/R)\phi_\alpha > e + (f(Q - d) + g(R - \beta))\phi_a = e + \xi\phi_a$, it follows that $r((-\Delta + p(e))^{-1}(e + f\bar{u} + g\bar{z} + p(e))) > 1$, and, therefore, that $e > \bar{e}$. Equality (4.3) implies that there exists $\varphi > 0$ such that $-\Delta\varphi = (e + \xi\phi_a)\varphi$, in Ω and $\varphi = 0$, on $\partial\Omega$. Arguing as in the proof of Lemma 1.4, we infer that $0 \leqq \int_\Omega [-\Delta\omega \cdot \omega - \omega^2(e + \xi\phi_a)]\, dx = \int_\Omega [\lambda_1\omega^2 - e\omega^2 - \xi\phi_a\omega^2]\, dx$. Therefore,

$$e \int_\Omega \omega^2\, dx \leqq \lambda_1 \int_\Omega \omega^2\, dx - \xi \int_\Omega \phi_a\omega^2\, dx. \tag{4.4}$$

On the other hand, $\phi_a \geqq \lambda_1(a - \lambda_1)\omega$ (see [CL, pp. 1128–1129]). This and (4.4) imply that

$$e \leqq \lambda_1 - \xi\lambda_1(a - \lambda_1) \int_\Omega \omega^3\, dx \Big/ \int_\Omega \omega^2\, dx.$$

Since $\bar{e} < e$, we get that $\bar{e} < \lambda_1(1 - \xi\kappa(a - \lambda_1))$, from which it follows that $\bar{e}$ is negative, for a big enough. □

REFERENCES

[A] H. AMANN, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[BB] J. BLAT AND K. J. BROWN, *Bifurcation of steady-state solutions in predator–prey and competition systems*, Proc. Roy. Soc. Edinburgh Sect. A, 97 (1984), pp. 21–34.

[CC] R. S. CANTRELL AND C. COSNER, *On the steady-state problem for the Volterra–Lotka competition model with diffusion*, preprint.

[CH] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. II, Interscience, New York, 1962.

[CL] 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.

[D1] E. N. DANCER, *On the indices of fixed points of mappings in cones and applications*, J. Math. Anal. Appl., 91 (1983), pp. 131–151.

[D2] ———, *On positive solutions of some pairs of differential equations*, Trans. Amer. Math. Soc., 284 (1984), pp. 729–743.

[D3] ———, *On positive solutions of some pairs of differential equations*, II, J. Differential Equation, 60 (1985), pp. 236–258.

[GT] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1977.

[HK] 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.

[K] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, 1966.

[KL] P. KORMAN AND A. LEUNG, *A general monotone scheme for elliptic systems with applications to ecological models*, Proc. Roy. Soc. Edinburgh Sect. A, 102 (1986), pp. 315–325.

[L1] A. LEUNG, *Equilibria and stabilities for competing species, reaction–diffusion equations with Dirichlet boundary data*, J. Math. Anal. Appl., 73 (1980), pp. 204–218.

[L2] ———, *Monotone schemes for semilinear elliptic systems related to ecology*, Math. Methods Appl. Sci., 4 (1982), pp. 272–285.

[MW] P. J. MCKENNA AND W. WALTER, *On the Dirichlet problem for elliptic systems*, Applicable Anal., 21 (1986), pp. 207–224.

[S] K. SCHMITT, *Boundary value problems for quasilinear second order elliptic equations*, Nonlinear Anal. Theory Methods Appl., 2 (1978), pp. 263–309.

[Sc] H. H. SCHAEFER, *Topological Vector-Spaces*, Springer-Verlag, Berlin, New York, 1971.

[W] G. T. WHYBURN, *Topological Analysis*, Princeton University Press, Princeton, NJ, 1958.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 660-672, May 1990
© 1990 Society for Industrial and Applied Mathematics
006

CONTACT MAPS AND INDUCED DIFFERENTIAL EQUATIONS*

HENRY HERMES†

Abstract. The goal of this paper is to examine constructive ways of generating solutions of partial differential equations (pde's) by the use of associated ordinary differential equations (ode's). Specifically, the classical method of Cauchy characteristics for first-order pde's is examined from the Cartan geometric viewpoint, i.e., as a method of extension via a contact (or Cauchy-characteristic) vector field on a submanifold determined by the equation in an appropriate jet bundle. Contact vector fields generate *contact transformations*, which are self maps of a jet bundle with induced cotangent space maps preserving the contact structure. *Contact maps*, or immersions, from one jet bundle to another with induced maps carrying contact structures in a specified manner are introduced. Contact transformations transform pde's to pde's; a contact map can induced an ordinary differential equation from a pde. For example, a contact map that induces a sixth-order, linear ode from the classical Burgers' pde is constructed. The final goal is to examine extending Cauchy data for second- or higher-order pde's via nonautonomous "Cauchy vector fields" on the jet bundles of the equations induced by contact maps.

Key words. contact transformations, Cauchy characteristics, geometric theory of pde's

AMS(MOS) subject classifications. 35A30, 58G37

Introduction. Our goal is to consider constructive ways of obtaining solutions of partial differential equations (pde's) by use of associated ordinary differential equations (ode's). The usual method of solving an initial value problem for a first-order pde is to extend the data along Cauchy characteristics, i.e., via the flow of a "Cauchy characteristic vector field" on the appropriate jet bundle associated with the equation. For a first-order equation, such a vector field can be explicitly constructed and will locally extend any (noncharacteristic) initial data to a solution. It is known that Cauchy characteristic vector fields do not exist on the jet bundles associated with pde's of order 2 or more. Our goal is to extend initial data for such equations via the flows arising from nonautonomous ode's on the associated jet bundle.

To be more specific, let $x=(x_1, x_2)\in\mathbb{R}^2$, $u^{(1)}=(u, u_{x_1}, u_{x_2})$ and $(x, u^{(1)})$ denote local coordinates on the jet bundle $J^1(\mathbb{R}^2, \mathbb{R}^1)$. Consider a first-order pde $\Delta(x, u^{(1)})=0$, i.e., $\Delta: J^1(\mathbb{R}^2, \mathbb{R}^1)\to\mathbb{R}^1$, and let M_Δ denote the manifold on which $\Delta=0$. Typical Cauchy data would be to prescribe a map $t\to x(t)\in\mathbb{R}^2$ and data $u(x(t))=u_0(t)$, $\partial u(x(t))/\partial x_1=p_1(t)$, $\partial u(x(t))/\partial x_2=p_2(t)$ where the compatibility conditions $u_0'(t)=p_1(t)x_1'(t)+p_2(t)x_2'(t)$ and $\Delta(x(t), u_0(t), p_1(t), p_2(t))=0$ must hold. In classical language this was expressed by calling $t\to v(t)=(x_1(t), x_2(t), u_0(t), p_1(t), p_2(t))\in M_\Delta$ an *initial strip*; in modern language v is called a *one-graph* in M_Δ. Geometrically, we picture v as defining a one-dimensional section of M_Δ. A parametric representation of a solution is obtained by extending the one-graph $t\to v(t)$ via the flow of Cauchy characteristics. Specifically, we construct a vector field V on $J^1(\mathbb{R}^2, \mathbb{R}^1)$, tangent to M_Δ, and with solution through q denoted $(\exp sV)(q)$, such that $\bar{v}(t, s)=(\exp sV)(v(t))$ is a one-graph giving a two-dimensional section of M_Δ from which the solution is obtained. In general, this geometric method of Cartan no longer works for higher-order equations $\Delta: J^k(\mathbb{R}^m, \mathbb{R}^1)\to\mathbb{R}^1$, $k\geqq 2$.

* Received by the editors May 31, 1989; accepted for publication (in revised form) July 17, 1989. This research was supported by National Science Foundation grant DMS-8500941 and by a visiting membership at the Mathematical Sciences Research Institute, Berkeley, California.

† Department of Mathematics, Box 426, University of Colorado, Boulder, Colorado 80309.

Section 1 introduces notation and summarizes basic results on k-graphs, contact transformations, and Cartan's geometrical viewpont of pde's. Section 2 introduces contact maps and their induced equations. Briefly, a kth-order pde in m independent variables, denoted $\Delta(x, u^{(k)}) = 0$, will be viewed as defining a submanifold M_Δ in the jet bundle $J^k(\mathbb{R}^m, \mathbb{R}^1)$. A contact transformation F is a self-diffeomorphism of $J^k(\mathbb{R}^m, \mathbb{R}^1)$ with induced cotangent space map F^* preserving the contact structure (i.e., if $\Omega^k(\mathbb{R}^m, \mathbb{R}^1)$ denotes the module of contact forms on $J^k(\mathbb{R}^m, \mathbb{R}^1)$, then $F^*\Omega^k = \Omega^k$). Form $F(M_\Delta)$, a new submanifold that may be viewed as defining an induced equation. Solutions of the latter are related to solutions of the former via F^{-1}. Example 1.3 will illustrate this classical use of contact transformations. A *contact map* F will be an immersion of one jet bundle $J^k(\mathbb{R}^m, \mathbb{R}^1)$ to another, say $J^l(\mathbb{R}^p, \mathbb{R}^q)$, $p \leqq m$, such that $F^*\Omega^l(\mathbb{R}^p, \mathbb{R}^q) \supset \Omega^k(\mathbb{R}^m, \mathbb{R}^1)$. Again, $F(M_\Delta)$ gives an induced equation, in particular, if $p = 1$ this will be an induced ordinary differential equation. If ψ is a solution of the induced equation, then F^{-1} of the graph of the lth prolongation of ψ (i.e., $\mathrm{pr}^{(l)}\,\psi$) gives a p-section (denote it $t \to v(t)$) of M_Δ, which is a k-graph. Thus there exist maps $x : \mathbb{R}^p \to \mathbb{R}^m$ and $w : \mathbb{R}^m \to \mathbb{R}^1$ such that $v(t) = (x(t), \mathrm{pr}^{(k)}\, w(x(t)))$ and $\Delta(x(t), \mathrm{pr}^{(k)}\, w(x(t))) = 0$. In this sense, w satisfies the pde Δ along the curve (submanifold) $t \to x(t)$. If $p = m-1$, i.e., $t = (t_1, \cdots, t_{m-1})$, then $t \to v(t)$ is a k-graph in M_Δ with image an $(m-1)$-dimensional section, and hence serves as Cauchy data for the pde $\Delta = 0$. The theory, to here, is illustrated by Example 2.2, which is Burgers' equation having $M_\Delta \subset J^2(\mathbb{R}^2, \mathbb{R}^1)$. We give a contact map $F : J^2(\mathbb{R}^2, \mathbb{R}^1) \to J^6(\mathbb{R}^1, \mathbb{R}^1)$ such that the induced equation is a linear, sixth-order, ordinary differential equation. Solutions of this linear ode readily provide appropriate Cauchy data for the pde.

Next, Theorem 1 gives conditions that, when satisfied, yield a vector field whose flow (when transverse) extends *any* k-graph $v : \mathbb{R}^q \to M_\Delta \subset J^k(\mathbb{R}^m, \mathbb{R}^n)$ to a k-graph $\bar{v} : \mathbb{R}^{q+1} \to M_\Delta$. As mentioned, if $k = 1$, this is classical, the vector field V is a contact vector field, indeed the Cauchy characteristic vector field. This is illustrated in Example 2.4. In general, the conditions of Theorem 1 cannot be satisfied for $k \geqq 2$. In § 3 we consider extensions via the flow of nonautonomous ode's. In particular, we consider the case $k = 2$ and a specific contact map $F : J^2(\mathbb{R}^2, \mathbb{R}^1) \to J^6(\mathbb{R}^1, \mathbb{R}^1)$. Here the induced equation associated with any second-order pde $\Delta : J^2(\mathbb{R}^2, \mathbb{R}^1) \to \mathbb{R}^1$ will be (in general) a sixth-order ode. We choose to do the extension in the jet bundle of the induced equation. If $\tau \to v(\tau) \in M_\Delta$ is a two-graph in $J^2(\mathbb{R}^2, \mathbb{R}^1)$ and $\gamma(\tau) = F \circ v(\tau)$, Theorem 2 gives conditions that, when satisfied, lead to a system of ode's on $J^6(\mathbb{R}^1, \mathbb{R}^1)$ such that their solution, denoted $s \to \phi(s, q)$, $\phi(0, q) = q$, satisfies $F^{-1} \circ \phi(s, \gamma(\tau))$ is a two-graph in M_Δ, which yields a parametric solution to the equation $\Delta = 0$ with initial data given by v.

Recent results in higher symmetries, conservation laws, and the Hamiltonian structure of partial differential equations have led to a rebirth and extension (in modern terminology) of the methods of Lie, Cartan [7], [5], and Vessiot [6]. These methods have also been instrumental in certain aspects of control theory [2], [3]. The rapid expansion of this subject has made much of the literature inaccessible to the non-specialist. Our goal here is to make the exposition as self-contained as possible (at the expense of generality) and to stress specific examples. Our notation is consistent with that in [5]. Certainly, a large part of this paper may be regarded as expository.

1. Graph maps and contact transformations. Local coordinates for a jet bundle $J^k(\mathbb{R}^m, \mathbb{R}^1)$ will be written as $(x, u^{(k)})$ where $x = (x_1, \cdots, x_m)$ and $u^{(k)} = (u, u_{x_1}, \cdots, u_{x_m}, u_{x_1x_1}, \cdots)$, i.e., subscripts to represent all partial derivatives through order k. As a specific example, local coordinates for $J^2(\mathbb{R}^2, \mathbb{R}^1)$ are $(x, u^{(2)}) =$

$(x_1, x_2, u, u_{x_1}, u_{x_2}, u_{x_1x_1}, u_{x_1x_2}, u_{x_2x_2})$. If $w:\mathbb{R}^m \to \mathbb{R}^1$ its kth prolongation, denoted $\mathrm{pr}^{(k)}w$, is defined as

$$\mathrm{pr}^{(k)}\, w(x) = (w(x), w_{x_1}(x), \cdots, w_{x_1x_1}(x), \cdots),$$

where here subscripts denote actual partial derivatives, and all partial derivatives through order k appear.

Given a differential equation $\Delta(x, u^{(k)}) = 0$, $x \in \mathbb{R}^m$, we consider $\Delta: J^k(\mathbb{R}^m, \mathbb{R}^1) \to \mathbb{R}^1$ and if this map has Jacobian of rank 1, then

$$M_\Delta = \{(x, u^{(k)}) \in J^k(\mathbb{R}^m, \mathbb{R}^1): \Delta(x, u^{(k)}) = 0\}$$

is a submanifold of codimension 1 in $J^k(\mathbb{R}^m, \mathbb{R}^1)$.

Example 1.1 (Burgers' equation).

$$\Delta(x, u^{(2)}) = u_{x_1} - uu_{x_2} - u_{x_2x_2}. \tag{1.1}$$

Here M_Δ is a seven-dimensional manifold in the eight-dimensional $J^2(\mathbb{R}^2, \mathbb{R}^1)$. A necessary and sufficient condition that a smooth function $w:\mathbb{R}^2 \to \mathbb{R}^1$ be a solution of $\Delta = 0$ is that the graph of $\mathrm{pr}^{(2)}\, w$ lies in M_Δ, i.e., $x \to (x, \mathrm{pr}^{(2)}\, w(x)) = (i, \mathrm{pr}^{(2)}\, w)(x) \in M_\Delta$. On the other hand, if $x \to (i, y)(x) = (x_1, x_2, y_1(x), \cdots, y_6(x)) \in M_\Delta$ we certainly cannot conclude that y_1 is a solution of $\Delta = 0$. Indeed, $x \to (i, y)(x) = (x_1, x_2, x_1^2 + x_2, 2x_1 + x_2, 1, 2, 0, 2x_1 - x_1^2) \in M_\Delta$ but clearly $y(x) \neq \mathrm{pr}^{(2)}\, w(x)$ for any $w:\mathbb{R}^2 \to \mathbb{R}^1$ since $\partial y_1(x)/\partial x_1 \neq y_2(x)$ as is necessary.

DEFINITION. For $p \leq m$, a map $v:\mathbb{R}^p \to J^k(\mathbb{R}^m, \mathbb{R}^1)$ is a *k-graph* if there exist maps $x:\mathbb{R}^p \to \mathbb{R}^m$, $w:\mathbb{R}^m \to \mathbb{R}^1$ such that $v(s) = (x(s), \mathrm{pr}^{(k)}\, w(x(s)))$, $s \in \mathbb{R}^p$. If x is an immersion, the image of v gives (locally) a p-dimensional section of $J^k(\mathbb{R}^m, \mathbb{R}^1)$.

Remark 1.1. The terminology k-graph stems from the standard case where $v:\mathbb{R}^m \to J^k(\mathbb{R}^m, \mathbb{R}^1)$ has the form $v(x) = (x, y(x)) = (i, y)(x)$ with the values $y(x)$ in the fibre above x. Here the image of v is the graph of y, and it is traditional to call y a k-graph if there exist $w:\mathbb{R}^m \to \mathbb{R}^1$ such that $y(x) = \mathrm{pr}^{(k)}\, w(x)$. This would be the nonparametric case: the above definition includes both the parametric case and a domain $\mathbb{R}^p$ with $p \leqq m$.

The conditions that partial derivatives match correctly in order that $v:\mathbb{R}^p \to J^k(\mathbb{R}^m, \mathbb{R}^1)$ be a k-graph are encoded in the module of contact forms, denoted $\Omega^k(\mathbb{R}^m, \mathbb{R}^1)$, on $j^k(\mathbb{R}^m, \mathbb{R}^1)$.

Example 1.2. The contact module $\Omega^2(\mathbb{R}^2, \mathbb{R}^1)$ is the module generated by the three one-forms:

$$\omega^1 = u_{x_1}\, dx_1 + u_{x_2}\, dx_2 - du, \qquad \omega^2 = u_{x_1x_1}\, dx_1 + u_{x_1x_2}\, dx_2 - du_{x_1},$$
$$\omega^3 = u_{x_1x_2}\, dx_1 + u_{x_2x_2}\, dx_2 - du_{x_2}$$

over the ring of smooth, real-valued, functions on $J^2(\mathbb{R}^2, \mathbb{R}^1)$.

PROPOSITION 1.1 (Gardner [1, Prop. 2.4]). *Let π denote projection of $J^k(\mathbb{R}^m, \mathbb{R}^1)$ onto $\mathbb{R}^m$, $p \leqq m$ and $v:\mathbb{R}^p \to J^k(\mathbb{R}^m, \mathbb{R}^1)$ be such that $\pi \circ v:\mathbb{R}^p \to \mathbb{R}^m$ is an immersion. Then a necessary and sufficient condition that v be a k-graph (locally) is that $v^*\Omega^k(\mathbb{R}^m, \mathbb{R}^1) = \{0\}$.*

The proof is exactly as that given by Gardner in [1] where he considered $v:\mathbb{R}^m \to J^k(\mathbb{R}^m, \mathbb{R}^1)$ with $\pi \circ v$ a diffeomorphism. This was used only to ensure that $\pi \circ v$ is locally one-to-one, hence his proof suffices for Proposition 1.1 as stated above.

Remark 1.2. In the nonparametric case where $v:\mathbb{R}^m \to J^k(\mathbb{R}^m, \mathbb{R}^1)$ has the form $v(x) = (i, y)(x)$ with $y(x)$ in the fibre above x, we have $\pi \circ v(x) = x$, i.e., the condition $\pi \circ v$, an immersion, is automatic.

In summary, and for convenience for later reference, we now have Proposition 1.2.

PROPOSITION 1.2. *Let $\Delta: J^k(\mathbb{R}^m, \mathbb{R}^1) \to \mathbb{R}^1$ be a given differential equation with Jacobian of rank 1 so M_Δ is well defined. If $v:\mathbb{R}^p \to M_\Delta \subset J^k(\mathbb{R}^m\mathbb{R}^1)$, $p \leqq m$, is such that*

$\pi \circ v$ is an immersion and $v^*\Omega^k(\mathbb{R}^m, \mathbb{R}^1) = \{0\}$, then there exist maps $x: \mathbb{R}^p \to \mathbb{R}^m$, $w: \mathbb{R}^m \to \mathbb{R}^1$ such that for $t \in \mathbb{R}^p$ (locally),

(a) $v(t) = (x(t), \operatorname{pr}^{(k)} w(x(t)))$.

(b) $\Delta(x(t), \operatorname{pr}^{(k)} w(x(t))) = 0$.

DEFINITION. A local diffeomorphism $F: J^k(\mathbb{R}^m, \mathbb{R}^1) \to J^k(\mathbb{R}^m, \mathbb{R}^1)$ is a *contact transformation* if F maps k-graphs into k-graphs, or equivalently, if its induced map on the cotangent bundle preserves the contact structure, i.e., $F^*\Omega^k(\mathbb{R}^m, \mathbb{R}^1) \subset \Omega^k(\mathbb{R}^m, \mathbb{R}^1)$.

Next we give a classical use of a contact transformation to "linearize" a nonlinear differential equation. The example is meant to motivate the concept of contact map and induced equation, which will follow.

Example 1.3. Consider $F: J^1(\mathbb{R}^1, \mathbb{R}^1) \to J^1(\mathbb{R}^1, \mathbb{R}^1)$ with (x, y, y') local coordinates in the domain and $(t, \sigma, \dot{\sigma})$ local coordinates in the range. Specifically, $t = F_1(x, y, y') = y'$, $\sigma = F_2(x, y, y') = y - xy'$, and $\dot{\sigma} = F_3(x, y, y') = -x$. Then F is a contact transformation (the Legendre transformation). To verify, $\Omega^1(\mathbb{R}^1, \mathbb{R}^1)$ is generated by $\omega(t, \sigma, \dot{\sigma}) = \dot{\sigma}dt - d\sigma$, hence (in local coordinates) $\omega(F^{-1}(t, \sigma, \dot{\sigma})) = (-x, -1, 0)$ and $F^*\omega = (-x, -1, 0)(\partial(F_1, F_2, F_3)/\partial(x, y, y')) = (y', -1, 0) = \omega(x, y, y')$ as required.

Next, consider the differential equation

$$\Delta(x, y, y') = x + \sum_{i=1}^{5} c_i (y')^i$$

where $c_1, \cdots, c_5$ are arbitrary constants. Here M_Δ is a two-manifold in $J^1(\mathbb{R}^1, \mathbb{R}^1)$, as is $F(M_\Delta)$. Indeed, the induced (or transformed) equation is

$$\tilde{\Delta}(t, \sigma, \dot{\sigma}) = -\dot{\sigma} + \sum_{i=1}^{5} c_i t^i$$

and $F(M_\Delta) = \{(t, \sigma, \dot{\sigma}): \tilde{\Delta} = 0\}$. Here, the induced equation is trivially integrated, i.e., $\psi(t) = k + \sum_{i=1}^{5} c_i t^{i+1}/(i+1)$ is its general solution. Then $t \to (i, \operatorname{pr}^{(1)} \psi)(t) \in F(M_\Delta)$ is a one-graph in $J^1(\mathbb{R}^1, \mathbb{R}^1)$. Its preimage $t \to v(t) = F^{-1}(t, \psi(t), \dot{\psi}(t)) = (-\dot{\psi}(t)), \psi(t) - t\dot{\psi}(t), t)$ is a one-section of M_Δ, which is a one-graph. Proposition 1.2 applies when $\dot{\psi}(t) \neq 0$, in which case we have the parametric representation of a solution of $\Delta = 0$ given by $x(t) = -\dot{\psi}(t)$, $y(t) = \psi(t) - t\dot{\psi}(t)$. Furthermore, with $t \to x(t)$ as above, Proposition 1.2 not only assures the existence of the nonparametric solution $w: \mathbb{R}^1 \to \mathbb{R}^1$ (in general difficult to find) but also gives the information that $w(x(t)) = \psi(t) - t\dot{\psi}(t)$, $w'(x(t)) = t$.

Remark 1.3. A coordinate change in $\mathbb{R}^m$ induces a contact transformation on $J^k(\mathbb{R}^m, \mathbb{R}^1)$ and, more generally, any self-diffeomorphism of $\mathbb{R}^m \times \mathbb{R}^1$ induces, by prolongation, a contact transformation of $J^k(\mathbb{R}^m, \mathbb{R}^1)$. (See [5] for the explicit construction.) Furthermore, a result of Bäcklund shows every contact transformation of $J^k(\mathbb{R}^m, \mathbb{R}^n)$ with $n \geqq 2$ is of this form, i.e., a prolongation of a self-diffeomorphism of $\mathbb{R}^m \times \mathbb{R}^n$ (see [7]). Two partial differential equations that can be transformed into each other via a local coordinate change in the underlying space should certainly be considered equivalent. A broader equivalence is that provided by the group of contact transformations in the appropriate jet bundle, a classification problem studied by Lie at the turn of the century.

2. Contact maps. Contact transformations are self-maps of a jet bundle, whereas contact maps may be from one jet bundle to another.

DEFINITION. An immersion $F: J^k(\mathbb{R}^m, \mathbb{R}^n) \to J^l(\mathbb{R}^p, \mathbb{R}^q)$ is a *contact map* if $F^*\Omega^l(\mathbb{R}^p, \mathbb{R}^q) \supset \Omega^k(\mathbb{R}^m, \mathbb{R}^n)$.

The motivation is as follows. Given a differential equation Δ with $M_\Delta \subset J^k(\mathbb{R}^m, \mathbb{R}^n)$ and contact map $F: J^k(\mathbb{R}^m, \mathbb{R}^n) \to J^l(\mathbb{R}^p, \mathbb{R}^q)$ we again have an induced equation

described by $F(M_\Delta)$. Suppose ψ is a solution of this induced equation, i.e., $t \to (i, \mathrm{pr}^{(l)}\psi)(t) \in F(M_\Delta)$. Let $v(t) = F^{-1}((i, \mathrm{pr}^{(l)}\psi)(t))$; so (locally) $t \to v(t) \in M_\Delta$ is a p-section of M_Δ. Then $(i, \mathrm{pr}^{(l)}\psi) = F \circ v$ and since $t \to (i, \mathrm{pr}^{(l)}\psi)(t)$ is an l-graph, $\{0\} = (i, \mathrm{pr}^{(l)}\psi)^*\Omega^l(\mathbb{R}^p, \mathbb{R}^q) = v^*F^*\Omega^l(\mathbb{R}^p, \mathbb{R}^q)$. But $F^*\Omega^l(\mathbb{R}^p, \mathbb{R}^q) \supset \Omega^k(\mathbb{R}^m, \mathbb{R}^n)$, hence $v^*\Omega^k = \{0\}$, i.e., $t \to v(t)$ is a k-graph in M_Δ to which we can, in general, apply Proposition 1.2.

The major portion of this section will deal with specific examples. Methods for the construction of contact transformations are well known, [7, §§ 2.3, 2.4]. No similar general results are known (at this time) for contact maps.

Our first goal is to construct a contact map $F: J^2(\mathbb{R}^2, \mathbb{R}^1) \to J^6(\mathbb{R}^1, \mathbb{R}^1)$, both jet bundles of dimension 8. When accomplished, the induced equation associated with *any* second-order pde in two independent variables, will be a sixth-order ode. The particular contact map we will exhibit is constructed to have a linear induced equation associated to the Burgers equation.

The generators of $\Omega^2(\mathbb{R}^2, \mathbb{R}^1)$, denoted ω^1, ω^2, ω^3, are exhibited in Example 1.2. Choose local coordinates for $J^6(\mathbb{R}^1, \mathbb{R}^1)$ as $(t, \sigma, \sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)}, \sigma^{(4)}, \sigma^{(5)}, \sigma^{(6)})$. Then generators of $\Omega^6(\mathbb{R}^1, \mathbb{R}^1)$ are:

$$\nu^1 = \sigma^{(1)}\,dt - d\sigma, \qquad \nu^2 = \sigma^{(2)}\,dt - d\sigma^{(1)},$$
$$\nu^3 = \sigma^{(3)}\,dt - d\sigma^{(2)}, \qquad \nu^4 = \sigma^{(4)}\,dt - d\sigma^{(3)},$$
$$\nu^5 = \sigma^{(5)}\,dt - s\sigma^{(4)}, \qquad \nu^6 = \sigma^{(6)}\,dt - d\sigma^{(5)}.$$

A direct attempt to find F such that $F^*\Omega^6 \supset \Omega^2$ leads to an extremely complicated set of partial differential equations for the components of F, aggravated by the fact that Ω^2 has three generators "to be covered." The construction is more tractable in the dual setting.

For X a vector field, say on $J^k(\mathbb{R}^m, \mathbb{R}^n)$, and ω a one-form, let $\langle \omega, X \rangle$ denote the standard, bilinear pairing of the cotangent bundle $T^*J^k(\mathbb{R}^m, \mathbb{R}^n)$ and tangent bundle $TJ^k(\mathbb{R}^m, \mathbb{R}^n)$. Dual to $\Omega^k(\mathbb{R}^m, \mathbb{R}^n)$ is a distribution (or module) sometimes called the Vesiot distribution:

$$\Lambda^k(\mathbb{R}^m, \mathbb{R}^n) = \{X \in TJ^k(\mathbb{R}^m, \mathbb{R}^n) \colon \langle \omega, X \rangle = 0, \qquad \omega \in \Omega^k(\mathbb{R}^m, \mathbb{R}^n)\}.$$

As before, F^* denotes an induced cotangent space map, whereas F_* will denote the induced tangent space map.

PROPOSITION 2.1. *Let $F: J^2(\mathbb{R}^2, \mathbb{R}^1) \to J^6(\mathbb{R}^1, \mathbb{R}^1)$ be an immersion and restrict attention to a neighborhood where F is one-to-one. Then $\Omega^2(\mathbb{R}^2, \mathbb{R}^1) \subset F^*\Omega^6(\mathbb{R}^1, \mathbb{R}^1)$ if and only if $\Lambda^6(\mathbb{R}^1, \mathbb{R}^1) \subset F_*\Lambda^2(\mathbb{R}^2, \mathbb{R}^1)$.*

Proof. (a) Suppose $\Omega^2 \subset F^*\Omega^6$ but $\Lambda^6 \not\subset F_*\Lambda^2$. Then there exist $X \in \Lambda^6$ and a one-form ν such that (i) $\langle \nu, X \rangle = 0$, and (ii) $\langle \nu, F_*\Lambda^2 \rangle = \{0\}$. But (i) implies $\nu \notin \Omega^6$, whereas (ii) yields $\langle F_*\nu, \Lambda^2 \rangle = \{0\}$ or $F^*\nu \in \Omega^2$. But F^* is one-to-one, hence $\nu \in \Omega^6$, a contradiction.

(b) Suppose $\Lambda^6 \subset F_*\Lambda^2$ but $\Omega^2 \not\subset F^*\Omega^6$. Then there exist $\omega \in \Omega^2$ and a vector field Y such that (i) $\langle \omega, Y \rangle \neq 0$, and (ii) $\langle F^*\Omega^6, Y \rangle = \{0\}$. But (ii) implies that $F_*Y \in \Lambda^6$ or $Y \in \Lambda^2$, which means $\langle \omega, Y \rangle = 0$ contradicting (i). □

As mentioned, the advantage of the use of Proposition 2.1 in our case is one of dimension, i.e., $\Lambda^6(\mathbb{R}^1, \mathbb{R}^1)$ has only two generators, specifically,

$$X^1 = \frac{\partial}{\partial x_1} + u\frac{\partial}{\partial x_2} + u_{x_1}\frac{\partial}{\partial u} + u_{x_2}\frac{\partial}{\partial u_{x_1}} + u_{x_1x_1}\frac{\partial}{\partial u_{x_2}} + u_{x_1x_2}\frac{\partial}{\partial u_{x_1x_1}} + u_{x_2x_2}\frac{\partial}{\partial u_{x_1x_2}},$$

$$X^2 = \frac{\partial}{\partial u_{x_2x_2}}.$$

The generators of $\Lambda^2(\mathbb{R}^2, \mathbb{R}^1)$ are

$$Y^1 = \frac{\partial}{\partial \sigma^{(4)}}, \quad Y^2 = \frac{\partial}{\partial \sigma^{(5)}}, \quad Y^3 = \frac{\partial}{\partial \sigma^{(6)}},$$

$$Y^4 = \frac{\partial}{\partial t} + \sigma^{(2)} \frac{\partial}{\partial \sigma^{(1)}} + \sigma^{(4)} \frac{\partial}{\partial \sigma^{(2)}} + \sigma^{(5)} \frac{\partial}{\partial \sigma^{(3)}},$$

$$Y^5 = \frac{\partial}{\partial \sigma} + \sigma^{(3)} \frac{\partial}{\partial \sigma^{(1)}} + \sigma^{(5)} \frac{\partial}{\partial \sigma^{(2)}} + \sigma^{(6)} \frac{\partial}{\partial \sigma^{(3)}}.$$

We seek a local diffeomorphism $F: J^2(\mathbb{R}^2, \mathbb{R}^1) \to J^6(\mathbb{R}^1, \mathbb{R}^1)$ such that for smooth functions $\alpha_1, \cdots, \alpha_5$ and $\beta_1, \cdots, \beta_5$

$$\sum_{i=1}^{5} \alpha_i F_* Y^i = X^1, \qquad \sum_{i=1}^{5} \beta_i F_* Y^i = X^2. \tag{2.1}$$

We next give a solution to (2.1).

Example 2.1. *A contact map* $F: J^2(\mathbb{R}^2, \mathbb{R}^1) \to J^6(\mathbb{R}^1, \mathbb{R}^1)$.

$$\begin{aligned} &t = F_1(x, u^{(2)}) = x_1, \quad \sigma = F_2 = -x_2, \quad \sigma^{(1)} = F_3 = u, \quad \sigma^{(2)} = F_4 = u_{x_1} - uu_{x_2}, \\ &\sigma^{(3)} = F_5 = u_{x_1x_1} - u_{x_1}u_{x_2} + uu_{x_2}^2 - 2uu_{x_1x_2} + u^2 u_{x_2x_2}, \\ &\sigma^{(4)} = F_6 = u_{x_2}, \quad \sigma^{(5)} = F_7 = u_{x_1x_2} - uu_{x_2x_2}, \quad \sigma^{(6)} = F_8 = u_{x_2x_2}. \end{aligned} \tag{2.2}$$

For completeness, the (irrelevant) values α_i, β_i needed to verify that (2.1) holds are

$$\begin{aligned} \alpha_1 &= -(u_{x_2}^2 - 2u_{x_1x_2} + 2uu_{x_2x_2})(u_{x_1} - uu_{x_2}) - u_{x_2}(u_{x_1x_1} - uu_{x_1x_2}) \\ &\quad + (u_{x_1} - 2uu_{x_2})(u_{x_1x_2} - uu_{x_2x_2}) + 2uu_{x_2x_2} + 2uu_{x_1}u_{x_2x_2} - 2u^2u_{x_2}u_{x_2x_2} + u_{x_2}, \\ \alpha_2 &= u_{x_1}u_{x_2x_2} - uu_{x_2}u_{x_2x_2} + u_{x_2x_2}, \quad \alpha_3 = 0, \quad \alpha_4 = 1, \quad \alpha_5 = u, \\ \beta_1 &= u^2, \quad \beta_2 = u, \quad \beta_3 = 1, \quad \beta_4 = 0, \quad \beta_5 = 0. \end{aligned}$$

$F^{-1}: J^6(\mathbb{R}^1, \mathbb{R}^1) \to J^6(\mathbb{R}^2, \mathbb{R}^1)$ is, locally,

$$\begin{aligned} &x_1 = t, \quad x_2 = -\sigma, \quad u = \sigma^{(1)}, \quad u_{x_1} = \sigma^{(2)} + \sigma^{(1)}\sigma^{(4)}, \quad u_{x_2} = \sigma^{(4)}, \\ &u_{x_1x_1} = \sigma^{(3)} + \sigma^{(4)}\sigma^{(2)} + 2\sigma^{(1)}\sigma^{(5)} - 1(\sigma^{(1)})^2\sigma^{(6)}, \\ &u_{x_1x_2} = \sigma^{(5)} + \sigma^{(6)}, \qquad u_{x_2x_2} = \sigma^{(6)}. \end{aligned} \tag{2.3}$$

Example 2.2 (Application of F, as above, to Burgers' equation). With $\Delta(x, u^{(2)}) = u_{x_1} - uu_{x_2} - u_{x_2x_2}$ the induced equation (using (2.2)) is

$$\sigma^{(6)}(t) - \sigma^{(2)}(t) = 0. \tag{2.4}$$

We may readily write a basis for the solution space of the linear induced equation. If ψ denotes any solution

$$t \to v(t) = F^{-1}((i, \operatorname{pr}^{(6)} \psi)(t)) \tag{2.5}$$

is a two-graph in M_Δ. Using F^{-1}, computing shows that $\pi \circ v(t) = (t, -\psi(t))$ and this gives an immersion in $\mathbb{R}^2$, hence Proposition 1.2 applies. In particular, we know that the map $x: \mathbb{R}^1 \to \mathbb{R}^2$ is $(x(t)) = (t, -\psi(t))$ and there exists $w: \mathbb{R}^2 \to \mathbb{R}^2$ such that $\Delta(x(t), \operatorname{pr}^{(2)} w(x(t))) = 0$. While w may be difficult to find, from F^{-1} and (2.5) we do know that $w(x(t)) = \psi^{(1)}(t)$, $w_{x_1}(x(t)) = \psi^{(2)}(t) + \psi^{(1)}(t)\psi^{(4)}(t)$, etc. In short, for any solution ψ of the induced equation (2.4), $v(t)$ as given in (2.5) defines appropriate "Cauchy data" for an initial value problem for the Burgers equation.

Example 2.3. *A contact map* $F: J^1(\mathbb{R}^2, \mathbb{R}^1) \to J^3(\mathbb{R}^1, \mathbb{R}^1)$. Choose $(x, u^{(1)}) = (x_1, x_2, u, u_{x_1}, u_{x_2})$ as local coordinates for $J^1(\mathbb{R}^2, \mathbb{R}^1)$ and $(t, \sigma, \sigma', \sigma'', \sigma''')$ as local coordinates for $J^3(\mathbb{R}^1, \mathbb{R}^1)$. Then $\Omega^1(\mathbb{R}^2, \mathbb{R}^1)$ is generated by $\omega = u_{x_1}\, dx_1 + u_{x_2}\, dx_2 - du$, whereas $\Omega^3(\mathbb{R}^1, \mathbb{R}^1)$ has $\nu^1 = \sigma' dt - d\sigma$, $\nu^2 = \sigma'' dt - d\sigma'$, $\nu^3 = \sigma''' dt - d\sigma''$ as generators. Since $\Omega^1(\mathbb{R}^2, \mathbb{R}^1)$ has but one generator, Proposition 2.1 should not be used, i.e., we proceed directly to seeking a map F and smooth functions α_1, α_2, α_3 such that $\omega = \sum_1^3 \alpha_i F^* \nu^i = F^*(\sum_1^3 \alpha_i \nu^i)$. Explicitly, if the components of F are denoted as $F_i(x, u^{(1)})$, then we require (in coordinate form)

$$\begin{aligned} &(\alpha_1 F_3 + \alpha_2 F_4 + \alpha_3 F_5, -\alpha_1, -\alpha_2, -\alpha_2, -\alpha_3, 0) \\ &\qquad (\partial(F_1, \cdots, F_5)/\partial(x, u^{(1)})) = (u_{x_1}, u_{x_2}, -1, 0, 0). \end{aligned} \tag{2.6}$$

A solution of (2.6) is

$$\begin{aligned} &t = F_1(x, u^{(1)}) = x_1, \quad \sigma = F_2 = x_2, \quad \sigma' = F_3 = u + u_{x_2}, \\ &\sigma'' = F_4 = -u_{x_2}, \qquad \sigma''' = F_5 = u_{x_1} + u_{x_2} + u u_{x_2} + u_{x_2}^2. \end{aligned} \tag{2.7}$$

The corresponding (irrelevant) values of the coefficient functions in (2.6) are $\alpha_1 = -u_{x_2}$, $\alpha_2 = \alpha_3 = 1$. An easy calculation gives F^{-1} as

$$x_1 = t, \quad x_2 = \sigma, \quad u = \sigma' + \sigma'', \quad u_{x_1} = \sigma''' + \sigma'' + \sigma'\sigma'', \quad u_{x_2} = -\sigma''. \tag{2.8}$$

We next address the problem of extending a k-graph $v: \mathbb{R}^p \to J^k(\mathbb{R}^m, \mathbb{R}^n)$, $p < m$, to a k-graph $\bar{v}: \mathbb{R}^m \to J^k(\mathbb{R}^m, \mathbb{R}^n)$, with the requirement that if the graph of v is in M_Δ, then this also is true for the graph of $\bar{v}$.

Extensions of *k*-graphs. For V a vector field on $J^k(\mathbb{R}^m, \mathbb{R}^n)$ we denote its flow map at time s, initiating from initial data q at $s = 0$, by $(\exp sV)(q)$ and for fixed s, let $(\exp sV): J^k(\mathbb{R}^m, \mathbb{R}^n) \to J^k(\mathbb{R}^m, \mathbb{R}^n)$ denote the flow-induced diffeomorphism. If ω is a one-form, $L_V\omega$ denotes its Lie derivative with respect to V. For generality, we now consider a system of several, say r, kth-order partial differential equations in m independent and n dependent variables, i.e., $\Delta = (\Delta_1, \cdots, \Delta_r): J^k(\mathbb{R}^m, \mathbb{R}^n) \to \mathbb{R}^r$.

DEFINITION. A vector field V on $J^k(\mathbb{R}^m, \mathbb{R}^n)$ is a *Cauchy characteristic vector field* for the pde system $\Delta: J^k(\mathbb{R}^m, \mathbb{R}^n) \to \mathbb{R}^r$ if

(a) $V\Delta = 0$ on M_Δ, i.e., V is tangent to M_Δ.

(b) $L_V \Omega^k(\mathbb{R}^m, \mathbb{R}^n) \subseteq \Omega^k(\mathbb{R}^m, \mathbb{R}^n)$ on M_Δ (i.e., V is a contact vector field, see Remark 2.1).

(c) $\langle \omega, V \rangle = 0$ on M_Δ for $\omega \in \Omega^k(\mathbb{R}^m, \mathbb{R}^n)$.

Properties show that V is a classical symmetry generator for Δ. Property (c) ensures (as will be shown) that for fixed $q \in M_\Delta$ the map $s \to (\exp sV)(q)$ is a k-graph. The reason for the terminology is that if $\Delta: J^1(\mathbb{R}^m, \mathbb{R}^1) \to \mathbb{R}^1$, i.e., is a first-order pde, a characteristic vector field V of Δ determines the classical characteristic equations.

THEOREM 1 (Extension via Cauchy characteristics). *Assume that* $\Delta: J^k(\mathbb{R}^m, \mathbb{R}^n) \to \mathbb{R}^r$ *defines a system of partial differential equations with Jacobian of rank r so* $M_\Delta = \{(x, u^{(k)}) \in J^k: \Delta = 0\}$ *is (locally) a submanifold. For* $t \in \mathbb{R}^p$ *let* $t \to v(t) \in M_\Delta$ *be a k-graph with image a p-section in* M_Δ, $p < m$, *and (again)* $\pi: J^k(\mathbb{R}^m, \mathbb{R}^n) \to \mathbb{R}^m$ *is projection. If V is a Cauchy characteristic vector field for* Δ *and if the map*

$$(t, s) \to \bar{v}(t, s) = (\exp sV)(v(t))$$

is such that $\pi \circ \bar{v}$ *is an immersion (roughly speaking the orbits of V are transverse to the graph of v), then* $\bar{v}$ *defines a* $(p+1)$ *section of* M_Δ, *which is a k-graph extension of V.*

Proof. Since V is tangent to M_Δ and the graph of v lies in M_Δ, it follows that $\bar{v}$ takes values in M_Δ. Also $\bar{v}(t, 0) = v(t)$ showing $\bar{v}$ extends v, whereas $\pi \circ v$, an immersion, ensures the image of $\bar{v}$ is a $(p+1)$-section of M_Δ when the image of v is a p-section.

Let $(\exp sV)^*$ denote the cotangent space map induced by the (fixed s) diffeomorphism $p \to (\exp sV)(p)$, let v_1 denote the map $t \to \bar{v}(t, s)$ with s fixed, and let v_2 denote the map $s \to \bar{v}(t, s)$ with t fixed. It suffices to show that $v_i^* \Omega^k(\mathbb{R}^m, \mathbb{R}^n) = \{0\}$, $i = 1, 2$.

First, $v_1^* \Omega^k = v^*(\exp sV)^* \Omega^k = v^* \Omega^k = \{0\}$, which uses the fact that V is a contact vector field, i.e., $(\exp sV)^* \Omega^k = \Omega^k$, and that v was given as a k-graph, i.e., $v^* \Omega^k = \{0\}$.

Finally, noting that $(\partial/\partial s)(\exp sV)(v(t)) = V((\exp sV)(v(t)))$, we have $v_2^* \Omega^k = \langle \Omega^k, V \rangle$. But $\bar{v}$, hence also v_2, takes values only on M_Δ and by (b), $\langle \Omega^k, V \rangle = \{0\}$ on M_Δ. Thus $\pi \circ \bar{v}$ is an immersion and $\bar{v}^* \Omega^k = \{0\}$. The conclusion now follows from Proposition 1.1. □

Remark 2.1. V is a *contact vector field* on $J^k(\mathbb{R}^m, \mathbb{R}^n)$ if for small $|s|$, $q \to (\exp sV)(q)$ is a contact transformation, i.e., $(\exp sV)^* \Omega^k \subset \Omega^k$ or equivalently that its Lie derivative satisfy $L_V \Omega^k \subseteq \Omega^k$. As will be shown in Example 2.4, an underlying key to the Cauchy characteristics for first-order pde's is the fact that the general form of a contact vector field on $J^1(\mathbb{R}^m, \mathbb{R}^1)$ is known. For $J^k(\mathbb{R}^m, \mathbb{R}^n)$, special contact vector fields, sometimes called banal vector fields (see [7]), can be constructed as prolongations of vector fields on $\mathbb{R}^m \times \mathbb{R}^n$. Specifically, let X be a vector field on $\mathbb{R}^m \times \mathbb{R}^n$. Then for small $|s|$, $(\exp sX)$ maps the graph of a function $w: \mathbb{R}^m \to \mathbb{R}^n$ to the graph of a function $\tilde{w}$ (possibly with restricted domain). The kth prolongation of X, denoted $\mathrm{pr}^{(k)} X$, is defined as the vector field on $J^k(\mathbb{R}^m, \mathbb{R}^n)$ such that $(\exp s(\mathrm{pr}^{(k)} X))$ takes the graph of $\mathrm{pr}^{(k)} w$ to the graph of $\mathrm{pr}^{(k)} \tilde{w}$. This shows that $\mathrm{pr}^{(k)} X$ is a (local) contact vector field, i.e., its flow maps k-graphs to k-graphs. For the explicit construction of $\mathrm{pr}^{(k)} X$ see [5]. These banal vector fields are too restrictive to have use in Theorem 1, even in the case $k = 1$. (See Remark 2.3.)

Example 2.4 (Classical Cauchy characteristics as an application of Theorem 1). Consider the pde $\Delta = 0$ where

$$\Delta(x, u^{(1)}) = u_{x_1} + u_{x_2}^2. \tag{2.9}$$

Clearly, the Jacobian of Δ, as a map from $J^1(\mathbb{R}^2, \mathbb{R}^1)$ to $\mathbb{R}^1$, has rank 1 so M_Δ is well defined. For purposes of illustration we first construct an initial data one-graph $t \to v(t) \in M_\Delta$ via the use of the induced equation generated by the contact map $F: J^1(\mathbb{R}^2, \mathbb{R}^1) \to J^3(\mathbb{R}^1, \mathbb{R}^1)$ given in Example 2.3. Since $\Delta = F_5 + F_3 F_4 + F_4 + F_4^2$, the induced equation is

$$\sigma''' + \sigma' \sigma'' + \sigma'' + (\sigma'')^2 = 0. \tag{2.10}$$

This equation is highly nonlinear, but it is an ordinary differential equation. We may verify, for example, that $\psi(t) = e^{-t}$ is a solution. Let ψ denote an arbitrary solution of (2.10). Then

$$v(t) = F^{-1}((i, \mathrm{pr}^{(3)} \psi)(t)) = (t, \psi(t), \psi'(t) + \psi''(t), -(\psi''(t))^2, -\psi''(t)) \tag{2.11}$$

and $t \to v(t)$ gives a one-section of $M_\Delta \subset J^1(\mathbb{R}^2, \mathbb{R}^1)$, which is a one-graph. In the classical terminology (e.g., [4, p. 39]), the image of v is called an *initial strip manifold* and it is well known that we can extend v, locally, to a solution of $\Delta = 0$ by the method of characteristics. The relationship between this and Theorem 1 follows.

We next extend v via Theorem 1. Let $f: J^1(\mathbb{R}^2, \mathbb{R}^1) \to \mathbb{R}^1$ be an arbitrary smooth function. The general form of a contact vector field on $J^1(\mathbb{R}^2, \mathbb{R}^1)$ is

$$V = -f_{u_{x_1}} \frac{\partial}{\partial x_1} - f_{u_{x_2}} \frac{\partial}{\partial x_2} + (f - u_{x_1} f_{u_{x_1}} - u_{x_2} f_{u_{x_2}}) \frac{\partial}{\partial u} + (f_{x_1} + u_{x_1} f_u) \frac{\partial}{\partial u_{x_1}} + (f_{x_2} + u_{x_2} f_u) \frac{\partial}{\partial u_{x_2}}.$$

(i) The condition that V be tangent to M_Δ, i.e., $V\Delta = 0$ on M_Δ, is

$$f_{x_1} + u_{x_1} f_u + 2u_{x_2}(f_{x_2} + u_{x_2} f_u) = 0 \quad \text{when } u_{x_1} = -u_{x_2}^2$$

or

$$f_{x_1} - 2f_{x_2}u_{x_2} - f_u u_{x_2}^2 = 0.$$

As in the calculation of symmetry generators (which is exactly this step; see [5]) this requires $f_{x_1} = 0$, $f_{x_2} = 0$, $f_u = 0$ on M_Δ, i.e., we will consider

$$f = f(u_{x_1}, u_{x_2}). \tag{2.12}$$

(ii) The condition $\langle \omega, V \rangle = 0$ on M_Δ requires

$$f = 0 \quad \text{on } M_\Delta. \tag{2.13}$$

Conditions (2.12), (2.13) are simultaneously satisfied by the choice $f = \Delta = u_{x_1} + u_{x_2}^2$ giving

$$V = -\frac{\partial}{\partial x_1} - 2u_{x_2}\frac{\partial}{\partial x_2} - u_{x_2}^2\frac{\partial}{\partial u}. \tag{2.14}$$

Finally, since V is a contact vector field, $L_V\Omega^2 \subset \Omega^2$ and V, as given by (2.14), is a Cauchy characteristic vector field.

Remark 2.2. The generalization of the above construction to $\Delta: J^1(\mathbb{R}^m, \mathbb{R}^1) \to \mathbb{R}^1$ is immediate, i.e., the known form of a Cauchy characteristic vector field for Δ is

$$V = -\sum_1^m \left(\frac{\partial \Delta}{\partial u_{x_i}}\right)\frac{\partial}{\partial u_{x_i}} - \left(\sum_1^m u_{x_i}\frac{\partial \Delta}{\partial u_{x_i}}\right)\frac{\partial}{\partial u} + \sum_1^m \left(\frac{\partial \Delta}{\partial x_i} + u_{x_i}\frac{\partial \Delta}{\partial u}\right)\frac{\partial}{\partial u_{x_i}}.$$

Continuing with Example 2.4, if ψ is a solution of the induced equation (e.g., $\psi(t) = e^{-t}$) and $v(t) = F^{-1} \circ \psi(t)$, then

$$\begin{aligned}\bar{v}(t, s) &= (\exp sV)(v(t)) \\ &= (t - s, \psi(t) + 2\psi''(t)s, \psi'(t) + \psi''(t) - (\psi''(t))^2 s, -(\psi''(t))^2, -\psi''(t))\end{aligned}$$

gives a two-section of M_Δ if the Jacobian $\partial(t - s, \psi(t) + 2\psi''(t)s)/\partial(t, s)$ is nonsingular, which is a two-graph extension of v. Proposition 1.2 applies; the map $x: \mathbb{R}^2 \to \mathbb{R}^2$ is explicitly given by $x(t, s) = (t - s, \psi(t) + 2\psi''(t)s)$, and we are assured the existence of $w: \mathbb{R}^2 \to \mathbb{R}^1$ such that $\bar{v}(t, s) = (x(t, s), \mathrm{pr}^{(1)} w(x(t, s)))$ with $\Delta(x(t, s), \mathrm{pr}^{(1)} w(x(t, s))) = 0$. Recovering w normally requires use of the Implicit Function Theorem.

The classical characteristic equations (e.g., see [4]) associated with $\Delta = u_{x_1} + u_{x_2}^2$ are (using $p_i = u_{x_i}$) $\dot{x}_i = \partial\Delta/\partial p_i$, $\dot{u} = \sum p_i\, \partial\Delta/\partial p_i$, $\dot{p}_i = -\partial\Delta/\partial x_i - (\partial\Delta/\partial u)p_i$, $i = 1, 2$ which, for Δ as above, gives rise to the vector field $-V$, with V as in (2.14). Thus, for a first-order equation, the extension of an initial strip manifold via Theorem 1 and Cauchy characteristics is the same.

Remark 2.3. The vector field V, given in (2.14), is not a banal contact vector field, i.e., the prolongation of a vector field on $\mathbb{R}^2 \times \mathbb{R}^1$. Indeed, any such prolongation would have coefficients of $\partial/\partial x_1$, $\partial/\partial x_2$, and $\partial/\partial u$, which are functions only of x_1, x_2, u. Thus even in the case of a first-order equation, banal contact vector fields do not suffice.

3. Extensions in the jet bundle of the induced equation. As illustrated, and as was classically known, Theorem 1 applies to first-order pde's. It was also known at the turn of the century that Cauchy characteristics (i.e., extensions via vector fields) did not work for higher than first-order equations. This does not imply, however, that it is impossible to extend an initial data "strip" (i.e., in the case $\Delta: J^2(\mathbb{R}^2, \mathbb{R}^1) \to \mathbb{R}^1$ a two-graph $\tau \to v(\tau) \in M_\Delta$) into a two graph giving a two-dimensional section of M_Δ to which Proposition 1.2 can be applied via the flow of a (nonautonomous) differential equation. We could attempt this in the jet bundle of the original equation. For the

sake of exposition, and simplicity of geometry, we choose to do this in the jet bundle of the induced equation and for the specific case of the contact map $F: J^2(\mathbb{R}^2, \mathbb{R}^1) \to J^6(\mathbb{R}^1, \mathbb{R}^1)$ as given in Example 2.1. Our setting for this section is, therefore, as follows:

(i) $F: J^2(\mathbb{R}^2, \mathbb{R}^1) \to J^6(\mathbb{R}^1, \mathbb{R}^1)$ as in Example 2.1.

(ii) $\Delta: J^2(\mathbb{R}^2, \mathbb{R}^1) \to \mathbb{R}^1$ is a second-order pde with Jacobian of maximal rank so $M_\Delta \subset J^2(\mathbb{R}^2, \mathbb{R}^1)$ is well defined (locally).

(iii) $\tilde{\Delta}: J^6(\mathbb{R}^1, \mathbb{R}^1) \to \mathbb{R}^2$ is the induced equation, assumed to have the form

$$\sigma^{(6)} - f(\sigma, \sigma^{(1)}, \cdots, \sigma^{(5)}) = 0. \tag{3.1}$$

(iv) $F(M_\Delta) = \{(t, \sigma, \cdots, \sigma^{(6)}) \in J^6(\mathbb{R}^1, \mathbb{R}^1): \tilde{\Delta} = 0\}$.

(v) $$\Omega_3^6 = F^{-1*}\Omega^2 \quad \text{so} \quad F^*\Omega_3^6 = \Omega^2(\mathbb{R}^2, \mathbb{R}^1). \tag{3.2}$$

Explicitly, Ω_3^6 is generated by

$$\mu^1 = (\sigma^{(2)} + \sigma^{(1)}\sigma^{(4)}, -\sigma^{(4)}, -1, 0, 0, 0, 0, 0),$$

$$\mu^2 = (\sigma^{(3)} + \sigma^{(4)}\sigma^{(2)} + 2\sigma^{(1)}\sigma^{(5)} - (\sigma^{(1)})^2\sigma^{(6)}, -\sigma^{(5)} - \sigma^{(1)}\sigma^{(6)}, -\sigma^{(4)}, -1, 0, -\sigma^{(1)}, 0, 0),$$

$$\mu^3 = (\sigma^{(5)} + \sigma^{(1)}\sigma^{(6)}, -\sigma^{(6)}, 0, 0, -1, 0, 0).$$

(vi) $\pi: J^2(\mathbb{R}^2, \mathbb{R}^1) \to \mathbb{R}^2$ is projection.

Vector fields whose flows extend arbitrary two-graphs $\tau \to v(\tau) \in M_\Delta$, and therefore vector fields whose flows extend arbitrary six-graphs $\tau \to \gamma(\tau) \in F(M_\Delta)$, cannot be expected. Should such an extending vector field exist, the properties it must satisfy are given in the next proposition (basically Theorem 1 restated in the jet bundle of the induced equation), which serves to motivate the properties required for an extension via a (nonautonomous) differential equation.

PROPOSITION 3.1 (Motivational). *Let V be a vector field on $J^6(\mathbb{R}^1, \mathbb{R}^1)$ and $\gamma: \mathbb{R}^1 \to F(M_\Delta) \subset J^6(\mathbb{R}^1, \mathbb{R}^1)$ be such that*:

(a) *V is tangent to $F(M_\Delta)$, i.e., $V\tilde{\Delta} = 0$ on $F(M_\Delta)$.*

(b) *$\langle \Omega_3^6, V \rangle = 0$ on $F(M_\Delta)$.*

(c) *$L_V\Omega_3^6 \subset \Omega_3^6$.*

(d) *$\gamma^*\Omega_3^6 = \{0\}$.*

Define

$$\bar{v}(\tau, s) = F^{-1} \circ (\exp sV) \circ \gamma(\tau). \tag{3.3}$$

Then $\bar{v}^*\Omega^2(\mathbb{R}^2, \mathbb{R}^1) = \{0\}$ and if $\pi \circ \bar{v}: \mathbb{R}^2 \to \mathbb{R}^2$ is an immersion, $(\tau, s) \to \bar{v}(\tau, s)$ is a two-graph in M_Δ to which Proposition 1.2 applies, i.e., $\bar{v}$ gives a parametric solution of $\Delta = 0$.

The proof of Proposition 3.1 is essentially the same as that for Theorem 1 and hence will be omitted.

Let D_t denote "total derivative," i.e., $D_t f(\sigma, \cdots, \sigma^{(5)}) = (\partial f/\partial\sigma)\sigma^{(1)} + \cdots + (\partial f/\partial\sigma^{(5)})\sigma^{(6)}$. We may readily show that through each point $q \in F(M_\Delta)$ there is a *unique* six-graph $s \to (\exp sW)(q) \in F(M_\Delta)$ where

$$W = \frac{\partial}{\partial t} + \sigma^{(1)}\frac{\partial}{\partial\sigma} + \cdots + \sigma^{(6)}\frac{\partial}{\partial\sigma^{(5)}} + (D_t f)\frac{\partial}{\partial\sigma^{(6)}}. \tag{3.4}$$

With this uniqueness in mind, the hope for Proposition 3.1 stems from the fact that although the maps $\tau \to \gamma(\tau) \in F(M_\Delta)$, $s \to (\exp sV)(\gamma(\tau)) \in F(M_\Delta)$ must be transverse (in order that $\pi \circ \bar{v}$ be an immersion) they need not be six-graphs. On the other hand, their images in M_Δ under F^{-1} are two-graphs, and we may verify that if V exists satisfying Proposition 3.1, then $F_*^{-1}V$ is an extending vector field on M_Δ for $\tau \to v(\tau) = F^{-1} \circ \gamma(\tau)$, which is not to be expected.

The method of extension is, therefore, via the flow of a (usually nonautonomous) ode on $J^6(\mathbb{R}^1, \mathbb{R}^1)$, and this ode will not extend arbitrary maps $\gamma: \mathbb{R}^1 \to F(M_\Delta)$ with $\gamma^* \Omega_3^6 = \{0\}$ but instead it will be constructed relative to a specific such γ.

3.1. Extensions via solutions of nonautonomous equations. It is convenient, in view of hypothesis (b) of Proposition 3.1, to have a basis for the vector fields V that satisfy $\langle \Omega_3^6, V\rangle = \{0\}$. We can, by inspection, find vector fields $W^1, \cdots, W^5$ on $J^2(\mathbb{R}^2, \mathbb{R}^1)$ that form a basis for the vector fields W such that $\langle \Omega^2, W\rangle = \{0\}$. Then since $F^{-1*}\Omega^2 = \Omega_3^6$, we have $\{0\} = \langle F^* F^{-1*} \Omega^2, W^i\rangle = \langle \Omega_3^6, F_* W^i\rangle$. Computing $F_* W^i$, $i = 1, \cdots, 5$ and simplifying via linear combinations yields a basis for the vector fields V that satisfy $\langle \Omega_3^6, V\rangle = \{0\}$ consisting of

$$V^1 = \frac{\partial}{\partial t} + \sigma^{(1)} \frac{\partial}{\partial \sigma} + \sigma^{(2)} \frac{\partial}{\partial \sigma^{(1)}} + \sigma^{(3)} \frac{\partial}{\partial \sigma^{(2)}} + \sigma^{(5)} \frac{\partial}{\partial \sigma^{(4)}},$$

$$V^2 = \frac{\partial}{\partial \sigma^{(3)}}, \quad V^3 = \frac{\partial}{\partial \sigma^{(5)}}, \quad V^4 = \frac{\partial}{\partial \sigma^{(6)}},$$

$$V^5 = -\frac{\partial}{\partial \sigma} + \sigma^{(4)} \frac{\partial}{\partial \sigma^{(1)}} + (\sigma^{(5)} - (\sigma^{(4)})^2) \frac{\partial}{\partial \sigma^{(2)}} + \sigma^{(6)} \frac{\partial}{\partial \sigma^{(4)}}.$$

Let $V = \sum_{i=1}^5 a_i V^i$. With $\tilde{\Delta}$ of the form given in (3.1), the requirement $V\tilde{\Delta} = 0$ implies that

$$\begin{aligned} a_4 = {} & (a_1 \sigma^{(1)} - a_5) \frac{\partial f}{\partial \sigma} + (a_1 \sigma^{(2)} + a_5 \sigma^{(4)}) \frac{\partial f}{\partial \sigma^{(1)}} + (a_1 \sigma^{(3)} + a_5 \sigma^{(5)} - a_5 (\sigma^{(4)})^2) \frac{\partial f}{\partial \sigma^{(2)}} \\ & + a_2 \frac{\partial f}{\partial \sigma^{(3)}} + (a_1 \sigma^{(5)} + a_5 \sigma^{(6)}) \frac{\partial f}{\partial \sigma^{(4)}} + a_3 \frac{\partial f}{\partial \sigma^{(5)}}. \end{aligned} \tag{3.5}$$

In summary, the most general V that satisfies both $\langle \Omega_3^6, V\rangle = \{0\}$ and $V\tilde{\Delta} = 0$ is given by

$$\begin{aligned} V = {} & a_1 \frac{\partial}{\partial t} + (a_1 \sigma^{(1)} - a_5) \frac{\partial}{\partial \sigma} + (a_1 \sigma^{(2)} + a_5 \sigma^{(4)}) \frac{\partial}{\partial \sigma^{(1)}} + (a_1 \sigma^{(3)} + a_5 \sigma^{(5)} - a_5 (\sigma^{(4)})^2) \frac{\partial}{\partial \sigma^{(2)}} \\ & + a_2 \frac{\partial}{\partial \sigma^{(3)}} + (a_1 \sigma^{(5)} + a_5 \sigma^{(6)}) \frac{\partial}{\partial \sigma^{(4)}} + a_3 \frac{\partial}{\partial \sigma^{(5)}} + a_4 \frac{\partial}{\partial \sigma^{(6)}} \end{aligned} \tag{3.6}$$

with a_4 given by (3.5), a_1, a_2, a_3, a_5 arbitrary. In particular, if these depend on the local coordinates $(t, \sigma, \cdots, \sigma^{(6)})$, then V is a vector field. We also allow, however, that $a_i = a_i(s, t, \sigma, \cdots, \sigma^{(6)})$, $i = 1, 2, 3, 5$ where s is a real variable in which case we may view these a_i as "control functions" and the differential equations associated with (3.6) will no longer be autonomous. If $a_1 \neq 0$, it can be considered a common factor and hence with a change of independent variable, here s, this factor could be made one. Thus we assume $a_1 = 1$ and (with little loss of generality) that $a_i = a_i(s, \sigma, \cdots, \sigma^{(6)})$, $i = 2, 3, 5$ are smooth in $\sigma, \cdots, \sigma^{(6)}$ for fixed s and measurable in s for fixed $\sigma, \cdots, \sigma^{(6)}$.

Let $\tau \to \gamma(\tau) \in F(M_\Delta)$ satisfy $\gamma^* \Omega_3^6 = \{0\}$. Then explicitly, from (3.6), the differential equations that will be used to extend this initial data are (with a_4 given by (3.5))

$$\begin{aligned} & t'(s) = 1, \quad t(0) = \gamma_1(\tau), \qquad \sigma'(s) = \sigma^{(1)} - a_5, \quad \sigma(0) = \gamma_2(\tau), \\ & \sigma^{(1)\prime}(s) = \sigma^{(2)} + a_5 \sigma^{(4)}, \quad \sigma^{(1)}(0) = \gamma_3(\tau), \\ & \sigma^{(2)\prime}(s) = \sigma^{(3)} + a_5 \sigma^{(4)} - a_5 (\sigma^{(4)})^2, \quad \sigma^{(2)}(0) = \gamma_4(\tau), \\ & \sigma^{(3)\prime}(s) = a_2, \quad \sigma^{(3)}(0) = \gamma_5(\tau), \qquad \sigma^{(4)\prime}(s) = \sigma^{(5)} + a_5 \sigma^{(6)}, \quad \sigma^{(4)}(0) = \gamma_6(\tau), \\ & \sigma^{(5)\prime}(s) = a_3, \quad \sigma^{(5)}(0) = \gamma_7(\tau), \qquad \sigma^{(6)\prime}(s) = a_4, \quad \sigma^{(6)}(0) = \gamma_8(\tau). \end{aligned} \tag{3.7}$$

Let $s \to \phi(s, \gamma(\tau), a)$ denote the solution of this initial value problem for (3.7) for some choice $a = (a_2, a_3, a_5)$ and $\phi^*(s, q, a)$ denote the cotangent space map induced by $q \to \phi(s, q, a)$. Then

$$\phi(s, \gamma(\tau), a) \in F(M_\Delta), \qquad \langle \phi'(s, \gamma(\tau), a), \Omega_3^6 \rangle = \{0\}. \tag{3.8}$$

This means that if we now define

$$\bar{v}(\tau, s) = F^{-1} \circ \phi(s, \gamma(\tau), a), \tag{3.9}$$

then $\bar{v}(\tau, s) \in M_\Delta$, and if v_1 is the map $\bar{v}$ with τ fixed, i.e., $s \to \bar{v}(\tau, s) = v_1(s)$, then $v_1^* \Omega^2 = \{0\}$. What remains is that if v_2 is the map $\bar{v}$ with s fixed, i.e., $\tau \to \bar{v}(\tau, s) = v_2(\tau)$, then we would need $v_2^* \Omega^2 = \{0\}$. Computing gives $v_2^* \Omega^2 = v_2^* F^* \Omega_3^6 = \gamma^* \phi^* \Omega_3^6$, which yields the desired condition $\phi^* \Omega_3^6 \subset \ker \gamma^*$. We summarize these calculations as Theorem 2.

THEOREM 2. *Let $\tau \to v(\tau) \in M_\Delta$ be a given two-graph and $\gamma(\tau) = F \circ v(\tau)$. (Note that this implies $\gamma^* \Omega_3^6 = \{0\}$.) Suppose, in system* (3.7), *that $a(s, \sigma, \cdots, \sigma^{(6)}) = (a_2, a_3, a_5)$ can be chosen so that the corresponding solution $\phi(s, \gamma(\tau), a) = (\phi_1(s, \gamma(\tau), a), \cdots, \phi_8(s, \gamma(\tau), a))$ satisfies the following*:

(i) *$\phi^*(s, \gamma(\tau), a) \Omega_3^6 \subset \ker \gamma^*$ along $\tau \to \gamma(\tau)$.*

(ii) *The Jacobian $\partial(\phi_1(s, \gamma(\tau), a), \phi_2(s, \gamma(\tau), a))/\partial(s, \tau)$ is nonsingular.*

Then $\bar{v}(\tau, s) = F^{-1} \circ \phi(s, \gamma(\tau), a)$ is a two-graph in M_Δ such that $\pi \circ \bar{v}: \mathbb{R}^2 \to \mathbb{R}^2$ is an immersion, and hence (by Proposition 1.2) *$\bar{v}$ gives a parametric solution to the problem $\Delta = 0$ with initial data given by v.*

Proof. The conclusion $\bar{v}^* \Omega^2 = \{0\}$ follows from the discussion preceding the statement of Theorem 2. The condition $\pi \circ \bar{v}$, an immersion, follows immediately from (ii) and the first two components of F. □

Remark 3.1. We may define the Lie derivative of a one-form ω with respect to a "time varying" vector field V as follows. Let $V = \sum_{i=1}^n a_i(s, x) \partial/\partial x_i$ (s denotes time) and $\omega = \sum_{i=1}^n b_i(x)\, dx_i$. Then

$$L_V \omega = \sum_{j=1}^n \left(Vb_j + \sum_{i=1}^n b_i \left(\frac{\partial a_i}{\partial s} + \frac{\partial a_i}{\partial x_j} \right) \right) dx_j.$$

Now in Theorem 2, the requirement that v be a two-graph implies (as noted) that $\Omega_3^6 \subset \ker \gamma^*$. We can therefore replace the (weak but difficult to verify) condition (i) by the "infinitesimal" condition

$$L_V(\ker \gamma^*) \subset \ker \gamma^*, \tag{i'}$$

which implies $\phi^*(s, \gamma(\tau), a) \ker \gamma^* \subset \ker \gamma^*$.

Computing ϕ^* means computing the fundamental solution matrix of the variational equation associated with system (3.7). We next give an example.

Example 3.1 (Burgers' equation). Here the induced equation is $\sigma^{(6)} = f(\sigma, \cdots, \sigma^{(5)}) = \sigma^{(2)}$. Consider the special choice of control

$$\begin{aligned} a_2(s, \sigma, \cdots, \sigma^{(6)}) &= \alpha_2(s)\sigma + \cdots + \alpha_8(s)\sigma^{(6)}, \\ a_3(s, \sigma, \cdots, \sigma^{(6)}) &= \beta_2(s)\sigma + \cdots + \beta_8(s)\sigma^{(6)}, \\ a_5 &= 0. \end{aligned} \tag{3.10}$$

The first equation in system (3.7) can always be explicitly solved, i.e., $\phi_1(s, \gamma(\tau), a) = s + \gamma_1(\tau)$. With the above choice of a_2, a_3, a_5, the remaining seven equations satisfy a linear system, hence the associated variational equation is again the same linear system. In this case, $\phi^*(s, \gamma(\tau), a)$ is independent of γ.

Subcase (trivial). Let $\tau \to v(\tau) = (0, -\tau, 1, 0, 0, 0, 0, 0)$. Then $\gamma(\tau) = F \circ v(\tau) = (0, \tau, 1, 0, 0, 0, 0, 0)$. Note that γ is not a six-graph. Choose $\alpha_2 = \cdots = \alpha_8 = \beta_2 = \cdots = \beta_8 = 0$ in (3.10). Then

$$\phi^*(s, \gamma(\tau), a) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & s & s^2 & s^3 & 0 & 0 & 0 \\ 0 & 0 & 1 & s & s^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & s & 0 \\ 0 & 0 & 0 & 0 & s & 0 & 0 & 1 \end{bmatrix}.$$

In components

$$\phi^*\mu^1 = (\sigma^{(2)} + \sigma^{(1)}\sigma^{(4)}, -\sigma^{(4)}, -s\sigma^{(4)} - 1, -s^2\sigma^{(4)} - s, -s^3\sigma^{(4)} - s^2, 0, 0),$$

$$\phi^*\mu^2 = (\sigma^{(3)} + \sigma^{(4)}\sigma^{(2)} + 2\sigma^{(1)}\sigma^{(5)} - (\sigma^{(1)})^2\sigma^{(6)}, -\sigma^{(5)} - \sigma^{(1)}\sigma^{(6)}, -s\sigma^{(5)} - s\sigma^{(1)}\sigma^{(6)} - \sigma^{(4)},$$
$$-s^2\sigma^{(5)} - s^2\sigma^{(1)}\sigma^{(6)} - s\sigma^{(4)} - 1, -s^3\sigma^{(5)} - s^3\sigma^{(1)}\sigma^{(6)} - s^2\sigma^{(4)} - s, -\sigma^{(1)}, -s\sigma^{(1)}, 0),$$

$$\phi^*\mu^3 = (\sigma^{(5)} + \sigma^{(1)}\sigma^{(6)}, -\sigma^{(6)}, -s\sigma^{(6)}, -s^2\sigma^{(6)}, -s^3\sigma^{(6)}, -1, -s, 0).$$

Here $\ker \gamma^*$ is spanned by dt, $d\sigma^{(1)}$, $d\sigma^{(2)}$, $d\sigma^{(3)}$, $d\sigma^{(4)}$, $d\sigma^{(5)}$, $s\sigma^{(6)}$ and along $\gamma(\tau)$, $\phi^*\mu^i \in \ker \gamma^*$, $i = 1, 2, 3$. Next, $\phi(s, \gamma(\tau), a) = (s, \tau + s, 1, 0, 0, 0, 0, 0)$ so the Jacobian condition (ii) of Theorem 2 is satisfied. Finally, $\bar{v}(\tau, s) = F^{-1} \circ \phi(s, \gamma(\tau), a) = (s, -\tau - s, 1, 0, 0, 0, 0, 0)$, which merely extends the constant data $u(x_1(\tau, 0), x_2(\tau, 0)) = u(0, -\tau) = 1$ to $u(s, -\tau - s) = 1$, which is a solution of Burgers' equation. The point here was to provide as simple as possible an example to illustrate the computations involved.

REFERENCES

[1] R. B. GARDNER, *A differential geometric generalization of characteristics*, Comm. Pure Appl. Math., 22 (1969), pp. 597-626.

[2] ———, *Differential geometric methods interfacing control theory*, in Differential Geometric Control Theory, R. Brockett, R. Millman, H. Sussmann, eds., Progress in Mathematics Vol. 27, Birkhauser, Boston, 1983, pp. 117-180.

[3] R. B. GARDNER AND W. F. SHADWICK, *Feedback equivalence for general control systems*, preprint.

[4] F. JOHN, *Partial Differential Equations*, Second edition, Springer-Verlag, Berlin, New York, 1975.

[5] P. J. OLVER, *Applications of Lie Groups to Differential Equations*, Springer-Verlag, Berlin, New York, 1986.

[6] P. J. VASSILIOU, *Coupled systems of nonlinear wave equations and finite dimensional Lie algebras* I, Acta Appl. Math., 8 (1987), pp. 107-148.

[7] A. M. VINOGRADOV, *Local symmetries and conservation laws*, Acta Appl. Math., 2 (1984), pp. 21-78.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 673-692, May 1990
© 1990 Society for Industrial and Applied Mathematics
007

QUASI CONVERGENCE AND STABILITY FOR STRONGLY ORDER-PRESERVING SEMIFLOWS*

HAL L. SMITH† AND HORST R. THIEME†

Abstract. Hirsch's results concerning quasi convergence of almost all trajectories of strongly monotone semiflows are derived under weaker assumptions adopted from Matano. The proofs are based on a sequential limit set trichotomy, which follows from the nonordering principle and the limit set dichotomy. The assumption excluding totally ordered arcs of equilibria, which is required for the set of asymptotically stable points to be dense, is verified for dynamical systems that are analytic on the state space.

Key words. quasi convergence, monotone dynamical system, strongly order-preserving semiflow, open dense set of quasiconvergent points, limit set dichotomy, nonordering principle, sequential limit set trichotomy, totally ordered continua of equilibria, global asymptotic stability

AMS(MOS) subject classifications. 34C35, 34G20, 47H20

1. Introduction. In a recent paper [8], Hirsch establishes that most orbits of a strongly monotone semiflow on a strongly ordered space X tend to the set E of the equilibria. Somewhat more precisely, Hirsch shows that the set of all $x \in X$ for which $\omega(x)$, the omega (positive) limit set of x, satisfies $\omega(x) \subset E$, is a "large" (open dense, residual, set of full measure) subset of X. Points x for which $\omega(x) \subset E$ are called *quasiconvergent* points; they are called *convergent* points if $\omega(x)$ consists of a single point of E. The theory of monotone semiflows, as developed by Hirsch, is both powerful and deep. Unfortunately, it can also be awkward to apply in some situations. One reason is that the state space for the dynamical system must be a *strongly* ordered Banach space, that is, one for which the nonnegative cone has nonempty interior. It is quite often the case that the "natural" state space for the dynamical system does not have this property. We can usually rectify this problem by finding a Banach subspace, imbedded in the natural space, which is strongly ordered. We can then apply the results of the theory to this smaller strongly ordered space. In the case where the semiflow maps points in the natural state space continuously into the smaller space, perhaps for suitably large times, we can bootstrap the results obtained for the dynamical system on the smaller space up to the natural state space. Even when all this can be done, it is so inconvenient that it is desirable to have an approach available that avoids these difficulties from the beginning. Second, Hirsch's results concerning stable and asymptotically stable points [8, §§ 8-10] require the semiflow to be restricted to an invariant, strongly ordered subset X_0 of the space X. This condition essentially forces X_0 to be an open subset of X and precludes immediate application of the results to such important examples as the case $X_0 = X_+$, the nonnegative cone in a Banach space X, or when X_0 is an order interval.

Matano [11], [12] has outlined a competing theory of monotone dynamical systems, parallel to Hirsch's, which does not require that the space be strongly ordered. The main results of Matano have appeared in conference proceedings [11], [12] without proofs. One of the main results of Matano's theory, as does that of Hirsch, provides sufficient conditions for "most" points to be quasiconvergent.

In this paper, we combine the ideas of Hirsch and Matano to obtain a theory that improves several of the results of both authors, while at the same time being conceptually

* Received by the editors February 13, 1989; accepted for publication (in revised form) July 17, 1989. The research of the first author was supported in part by National Science Foundation grant DMS 8722279.
† Department of Mathematics, Arizona State University, Tempe, Arizona 85287.

simpler. We adopt (a slight generalization of) Matano's idea of a *strongly order-preserving* semiflow that does not require the space to be strongly ordered. On the other hand, all our results are based on modified versions of two fundamental results due to Hirsch, namely, the *nonordering principle* for limit sets [8] and the *limit set dichotomy* [8]. These two principles also hold under Matano's weaker assumptions (see [12]) and can actually be shown by modifying Hirsch's proofs accordingly. They alone lead to a simple proof of our Proposition 3.1 (*sequential limit set trichotomy*), a result from which most of our theory follows. In this sense, we feel that the theory develops more naturally and more simply.

In the remainder of this section we describe some basic ideas and notation and preview several of the main results.

Let X be an ordered metric space with metric d and order relation $\leqq$. We write $x < y$ if $x \leqq y$ and $x \neq y$. Points x and y in X are *ordered* if either $x < y$ or $y < x$. Given two subsets A and B of X we write $A \leqq B$ ($A < B$) whenever $x \leqq y$ ($x < y$) for each choice of $x \in A$ and $y \in B$. If $x < y$ then $[x, y] = \{z \in X : x \leqq z \leqq y\}$. A subset Y of X is *order convex* if $[y_1, y_2] \subset Y$ whenever $y_1, y_2 \in Y$ and $y_1 < y_2$.

We assume that the order and the topology on X are compatible in the sense that $x \leqq y$ whenever $x_n \to x$, $y_n \to y$, and $x_n \leqq y_n$ for all n. If $x \in X$ we say that x can be approximated from below (above) in X if there exists a sequence $\{x_n\}$ in X satisfying $x_n < x_{n+1} < x$ ($x < x_{n+1} < x_n$) for $n \geqq 1$ and $x_n \to x$.

Let $\Phi : X \times \Re^+ \to X$ be a semiflow on X, that is, Φ is continuous and $\Phi_t(x) \equiv \Phi(x, t)$ satisfies $\Phi_0(x) = x$ for every x and $\Phi_t \Phi_s = \Phi_{t+s}$ for every $t, s \geqq 0$. For $x \in X$, let

$$\mathcal{O}^+(x) = \{\Phi_t(x) : t \geqq 0\}, \qquad \omega(x) = \bigcap_{t \geqq 0} \overline{\mathcal{O}^+(\Phi_t(x))}$$

be the orbit initiating at x and the omega (positive) limit set of $\mathcal{O}^+(x)$, respectively. Of course the latter may be empty. Hence we require some compactness properties to hold for the semiflow Φ. Namely,

(C) For each $x \in X$, $\mathcal{O}^+(x)$ has compact closure in X. In addition, for each compact subset K of X, $\bigcup_{x \in K} \omega(x)$ has compact closure in X.

We will assume that (C) holds throughout the remainder of this section. It is a relatively mild compactness assumption, weaker than required by Matano [11], [12].

If $\mathcal{O}^+(x)$ has compact closure in X then $\omega(x)$ is nonempty, compact, connected, and invariant, i.e., $\Phi_t(\omega(x)) = \omega(x)$, $t \geqq 0$, and $\Phi_t(x) \to \omega(x)$ as $t \to \infty$. We let $E = \{x \in X : \Phi_t(x) = x, t \geqq 0\}$ be the set of equilibria. The set of quasiconvergent points is denoted by $Q = \{x \in X : \omega(x) \subset E\}$ and the set of convergent points by $C = \{X \in Q : \omega(x)$ is a singleton set$\}$.

The semiflow Φ is said to be *monotone* provided

$$\Phi_t(x) \leqq \Phi_t(y) \quad \text{whenever } x \leqq y.$$

Following Matano [11], [12], Φ is said to be *strongly order-preserving* if Φ is monotone, and whenever $x, y \in X$ with $x < y$, there exist open sets U and V, $x \in U$, $y \in V$, and $t_0 \geqq 0$ such that $\Phi_{t_0}(U) \leqq \Phi_{t_0}(V)$. By monotonicity, it follows that

$$\Phi_t(U) \leqq \Phi_t(V) \quad \text{for } t \geqq t_0.$$

One of our main results is the following theorem (cf. [12, Thm. 5] and [8, Thm. 7.5]).

THEOREM 1. *Let X be an ordered metric space and let Φ_t be a strongly order-preserving semiflow on X. Suppose that each point of X can be approximated from above or from below in X. Then* Int Q *is dense in X.*

Some remarks on the theorem are appropriate here. First, note that we require minimal assumptions on the space X. Since subsets of ordered metrizable spaces are themselves ordered metrizable spaces under the inherited order and metric, we may view X as a subset of some larger space. The requirement that every point may be approximated from above or from below in the space X is not a particularly strong one. For example, if X is the nonnegative cone or a nontrivial order interval in an ordered Banach space then X has the required property.

Theorem 1 is a generalized and simplified version of a result by Matano [12, Thm. 5]. In particular we do not require the existence of a bounded set that every orbit must eventually enter and remain in, and our compactness requirement (C) is milder than Matano's (H.2). Hirsch proves [8, Thm. 7.5 and Cor. 7.6] that Q (not Int Q) is dense without compactness hypotheses but with much stronger assumptions on the space. For smooth strongly monotone flows defined by semilinear parabolic equations, Poláčik [14] has shown that Int C is dense in X.

With an additional compatibility relation between the metric and the order, a significantly stronger conclusion can be drawn. The space X is said to be normally ordered if there exists a constant $k>0$ such that

$$d(u, v) \leqq k d(x, y)$$

for all x, y, u, v with $u, v \in [x, y]$. A subset X of an ordered Banach space possessing a normal positive cone (e.g., $C(\Omega)$, $L^p(\Omega)$) is normally ordered.

THEOREM 2. *Let X be a normally ordered metric space, and let Φ_t be a strongly order-preserving semiflow on X. Suppose that each point x in X can be approximated from above or from below in X, and there exists an open and dense subset X_0 in X with the property that each point of X_0 can be approximated both from above and from below in X. Then $A \cup \operatorname{Int}(S \cap C)$ is dense in X.*

The set A in Theorem 2 is the set of *asymptotically stable* points, as defined by Hirsch [8]. A point x belongs to A if there is a neighborhood V of x with the property that for every $\varepsilon>0$ there exists $t_\varepsilon>0$ such that $d(\Phi_t(x), \Phi_t(y))<\varepsilon$ if $t \geqq t_\varepsilon$ and $y \in V$. The set S is the set of stable points. A point x belongs to S if for every $\varepsilon>0$ there exists $\delta>0$ such that $d(\Phi_t(x), \Phi_t(y))<\varepsilon$ for $t \geqq 0$ whenever $y \in X$ and $d(x, y)<\delta$. Hirsch observes [8] (see also Proposition 3.4 in § 3) that A is open and $A \subset S \subset Q$. Note that the map $x \mapsto \omega(x)$ is locally constant near an asymptotically stable point and continuous at a stable point. Theorem 2 asserts that an open and dense set of points of X consists of either asymptotically stable quasiconvergent points or stable convergent points having the property that sufficiently nearby points are also convergent.

Theorem 2 may be compared with Proposition 9.5 of Hirsch [8], which draws essentially the same conclusion assuming the "astrictive" condition [8, Def. 8.9] and, of course, that X is strongly ordered and Φ is strongly monotone. The astrictive condition of Hirsch is difficult to verify directly. Hirsch provides several sufficient conditions [8, Thms. 8.11–8.14] for the astrictive condition to hold. Those sufficient conditions that are most likely to be useful in infinite-dimensional settings require a stronger compactness condition than (C).

Our next result improves the global stability result of Hirsch [8, Thm. 10.3] at the expense of an additional compactness condition (the second part of (C)).

THEOREM 3. *Let X be a connected, ordered metric space with the property that each point of X can be approximated from above and from below in X. If X does not contain two ordered equilibria, then X contains a unique equilibrium point to which every orbit converges.*

The assumption that X does not contain two equilibria $p, q \in E$, with $p < q$ certainly holds if X is known to contain at most one equilibrium point. Hirsch assumes the existence of a unique equilibrium point.

The following convergence result seems to be new.

THEOREM 4. *Suppose that X is an order convex subset of a normally ordered Banach space Y. Let Φ_t be a strongly order-preserving semiflow on X such that Φ_t is condensing on every order interval in X for $t > 0$. Assume that any point in $X \setminus E$ can be approximated both from above and below. If X contains at most two order-related equilibria, then $X = Q$. If X contains no more than two equilibria, then every trajectory in X converges towards one of them.*

This result is optimal insofar as five-dimensional single-loop positive-feedback systems (which induce strongly order-preserving semiflows) can have three order-related equilibria with the middle one undergoing an unstable Hopf bifurcation. See Selgrade [17].

Our next result provides sufficient conditions (similar to those in Hirsch [8, Thm. 9.6]) for the set A of asymptotically stable points to be dense in X. We need to assume that each point $x \in X$ belongs to some totally ordered arc in X. A subset J of X is a totally ordered arc provided $J = \psi(I)$ where I is a nontrivial interval of real numbers and ψ is a continuous function from I onto J such that $\psi(s) < \psi(t)$ whenever $s, t \in I$ and $s < t$.

THEOREM 5. *Suppose that X is a normally ordered metric space and that each $x \in X$ belongs to some totally ordered arc in X. Let Φ_t be a strongly order-preserving semiflow on X. If there does not exist a totally ordered arc of equilibria in X then A is open and dense in X.*

Section 4 is devoted to providing a verifiable sufficient condition for excluding the existence of totally ordered arcs of equilibria contained in bounded open subsets of a Banach lattice. By exploiting an analyticity hypothesis on the semiflow Φ, we are able to show that a bounded open set U contains no totally ordered arc of equilibria provided the boundary of U does not contain equilibria for which the infinitesimal generator of the variational semigroup has vanishing spectral bound (see Theorem 4.1). This result provides a testable condition in order that the main hypothesis of Theorem 4 holds.

For applications of the results of monotone dynamical systems theory, we refer the reader to [5]–[7], [18] for applications to systems of ordinary differential equations, to [8], [11], [12], [14], [15] for applications to parabolic initial boundary value problems, to [19], [20] for applications to functional differential equations, and to [9], [10] for applications to systems of parabolic equations with time delays. We conclude this section by giving a concrete example of a semiflow generated by a semilinear parabolic equation to which our results apply. We do not aim for utmost generality. Articles [8], [11], [12] contain applications to parabolic initial boundary value problems of much greater generality. Our example can even be treated by energy methods as it generates a gradient flow [4]. Our results, however, apply directly to a state space that is not strongly ordered. As it parallels a similar example treated in [8], we closely follow the treatment given there.

Let Ω be a bounded open set in $\mathbf{R}^n$ with boundary, $\partial\Omega$, of class $C^{2+\alpha}$, $\alpha > 0$. Let $f: \bar{\Omega} \times \mathbf{R} \to \mathbf{R}$ satisfy the following:

(F) (a) $f(x, 0) = 0$ for $x \in \partial\Omega$.

(b) There exist $a, b \in \mathbf{R}$, $a < b$, $a \leqq 0 \leqq b$, such that

$$f(x, b) \leqq 0 \leqq f(x, a), \qquad x \in \bar{\Omega}.$$

(c) f is a Lipschitz on $\bar{\Omega}\times[a, b]$ (in both arguments jointly).

Consider the initial boundary value problem

$$\text{(P)}\qquad \begin{aligned} &\frac{\partial u}{\partial t}=\Delta u+f(x, u), && x\in\Omega, \quad t>0,\\ &u=0, && x\in\partial\Omega, \quad t>0,\\ &u(x, 0)=u_0(x), && x\in\Omega. \end{aligned}$$

The Laplace operator Δ appearing in (P) could be replaced by any second-order uniformly strongly elliptic differential operator. The nonlinear term f could depend on the gradient of u with appropriate hypotheses [8], [11], [12]. Other boundary conditions could be considered, however, the Dirichlet boundary conditions will allow us to generate a strongly order-preserving semiflow on a natural state space, which is not a strongly monotone semiflow.

Let $C^r(\bar{\Omega})$, $r=0$, 1, be the Banach spaces of continuous ($r=0$) and continuously differentiable ($r=1$) functions on $\bar{\Omega}$ with the usual norms. Denote by ${}_0C^r(\bar{\Omega})$ the Banach subspace of functions vanishing on the boundary of Ω and by ${}_0C^r_+(\bar{\Omega})$ the cone of nonnegative functions in ${}_0C^r(\bar{\Omega})$. As noted in [8], $\operatorname{Int}{}_0C^0_+(\bar{\Omega})$ is empty while

$$\operatorname{Int}{}_0C^1_+(\bar{\Omega})=\left\{u\in{}_0C^1_+(\bar{\Omega}): \frac{\partial u}{\partial \nu}<0 \text{ on } \partial\Omega\right\}\neq\varnothing,$$

where ν is the outer normal unit vector field on $\partial\Omega$ and $\partial u/\partial\nu$ is the directional derivative in the direction ν. In particular, ${}_0C^0(\bar{\Omega})$ is not a strongly ordered space.

As in Hirsch, we observe the following facts. Mild solutions of (P) generate a local semiflow Φ on ${}_0C^0(\bar{\Omega})$ [8, Thm. 4.4], which satisfies the following:

(S) $\Phi_t: {}_0C^0(\bar{\Omega})\to{}_0C^1(\bar{\Omega})$ is continuous for $t>0$, and

(M) $u_0<u_1$ in ${}_0C^0(\bar{\Omega})$ implies $\Phi_t(u_0)<\Phi_t(u_1)$.

Fact (M) follows from the maximum principle and the fact that mild solutions are classical solutions for $t>0$ (see Remark 4.2 of [8]).

Let

$$X=\{u\in{}_0C^0(\bar{\Omega}): a\leqq u(x)\leqq b, x\in\bar{\Omega}\}.$$

X is a normally ordered metric space with the order and metric it inherits from ${}_0C^0(\bar{\Omega})$. Indeed, it is an order convex subset of the normally ordered Banach space ${}_0C^0(\bar{\Omega})$. It is easy to see that every point of X can be approximated from above or from below in X. Arguments similar to those yielding (M) give the following:

(I) Φ restricts to a monotone (global) semiflow on X.

Moreover, we have [8, Thm. 4.4] the following:

(K) Φ_t is completely continuous on X.

Our compactness hypothesis (C) follows immediately from (K).

Although X is not a strongly ordered space, since $\operatorname{Int}{}_0C^0(\bar{\Omega})=\varnothing$, and thus $\Phi_t: X\to X$ cannot be strongly monotone, we nonetheless have the following proposition.

PROPOSITION 6. *Φ is strongly order preserving on X.*

Proof. Let u_0, $v_0\in X$ with $u_0<v_0$. By (M), $\Phi_t(u_0)<\Phi_t(v_0)$ for $t>0$. Let $t_0>0$. Now $u(x, t)\equiv[\Phi_t(u_0)](x)$ is a classical solution ($t>0$) so maximum principle arguments apply. If there exists $x_0\in\Omega$ such that $v(x_0, t_0)=u(x_0, t_0)$ then $v(x, t)\equiv u(x, t)$

for $x \in \Omega$, $t \leqq t_0$ by the maximum principle. It follows that $v(x, t) > u(x, t)$ for all $x \in \Omega$, $t > 0$. But then the boundary point principle implies $\partial/\partial\nu(u - v) > 0$ on $\partial\Omega$. Thus

$$\Phi_{t_0}(v_0) - \Phi_{t_0}(u_0) \in \operatorname{Int}{}_0C^1_+(\bar{\Omega}).$$

Hence we may find neighborhoods U_0 and V_0 of $\Phi_{t_0}(u_0)$, $\Phi_{t_0}(v_0)$ in ${}_0C^1(\bar{\Omega})$ such that

$$U_0 \leqq V_0 \quad \text{in } {}_0C^1(\bar{\Omega}).$$

Since $\Phi_{t_0}: X \to {}_0C^1(\bar{\Omega})$ is continuous there are neighborhoods U and V of u_0 and v_0 in X such that

$$\Phi_{t_0}(U) \subset U_0, \qquad \Phi_{t_0}(V) \subset V_0.$$

Thus $\Phi_{t_0}(U) \leqq \Phi_{t_0}(V)$. This completes the proof.

The equilibria of Φ are given by the solutions $u \in X$ of the boundary value problem

$$\text{(E)} \qquad \begin{aligned} \Delta u + f(x, u) &= 0, \qquad x \in \Omega, \\ u &= 0, \qquad x \in \partial\Omega. \end{aligned}$$

Theorem 2 applies to the above example, implying that an open and dense set of initial conditions in X consists of asymptotically stable quasiconvergent points or stable convergent points. For these initial data, corresponding solutions converge to the set of solutions of (E) above.

In the following section we state some preliminary results. Section 3 contains the statements and proofs of our main results.

2. Nonordering principle and limit set dichotomy. Throughout this section we assume that X is an ordered metric space with metric d and that Φ is a semiflow on X, which is strongly order preserving. In addition, we assume that every orbit has compact closure in X. The following result is used extensively in this section.

LEMMA 2.1. *Let K_1 and K_2 be two compact subsets of X satisfying $K_1 < K_2$. Then there exist open sets U and V, $K_1 \subset U$, $K_2 \subset V$, and $t_1 \geqq 0$, $\varepsilon > 0$ such that*

$$\Phi_{t+s}(U) \leqq \Phi_t(V) \quad \textit{for } t \geqq t_1, \quad 0 \leqq s \leqq \varepsilon.$$

Proof. Let $x \in K_1$. For each $y \in K_2$ there exists $t_y \geqq 0$ and a neighborhood U_y of x and a neighborhood V_y of y such that $\Phi_t(U_y) \leqq \Phi_t(V_y)$ for $t \geqq t_y$ since Φ is strongly order preserving. Then $\{V_y\}_{y \in K_2}$ is an open cover of K_2, so we may choose a finite subcover: $K_2 \subset \bigcup_{i=1}^n V_{y_i} \equiv \tilde{V}$ where $y_i \in K_2$, $1 \leqq i \leqq n$. Let $\tilde{U} = \bigcap_{i=1}^n U_{y_i}$, so $\tilde{U}$ is a neighborhood of x and let $\tilde{t} = \max_{1 \leqq i \leqq n} t_{y_i}$. Then, for each i, $\Phi_{\tilde{t}}(\tilde{U}) \subset \Phi_{\tilde{t}}(U_{y_i}) \leqq \Phi_{\tilde{t}}(V_{y_i})$, so $\Phi_t(\tilde{U}) \leqq \Phi_t(V_{y_i})$ for each i, $1 \leqq i \leqq n$, and $t \geqq \tilde{t}$. It follows that $\Phi_t(\tilde{U}) \leqq \Phi_t(\tilde{V})$ for $t \geqq \tilde{t}$. For the remainder of the proof we write $\tilde{U}_x = \tilde{U}$ and $\tilde{V}_x = \tilde{V}$ to emphasize the dependence of these open sets on the point $x \in K_1$. Similarly $\tilde{t} = \tilde{t}_x$. Since $x \in K_1$ was arbitrary, we find for each $x \in K_1$ an open neighborhood $\tilde{U}_x$ of x, an open neighborhood $\tilde{V}_x$ of K_2, and $\tilde{t}_x \geqq 0$ such that $\Phi_t(\tilde{U}_x) \leqq \Phi_t(\tilde{V}_x)$ for $t \geqq t_x$. Again, $\{\tilde{U}_x\}_{x \in K_1}$ is an open cover of K_1, so we extract a finite subcover $\tilde{U}_{x_i}$, $1 \leqq i \leqq m$, and set $U = \bigcup_{i=1}^m \tilde{U}_{x_i} \supset K_1$, $V = \bigcap_{i=1}^m \tilde{V}_{x_i} \supset K_2$ and $t_1 = \max_{1 \leqq i \leqq m} \tilde{t}_{x_i}$. Since $V \subset \tilde{V}_{x_i}$, $\Phi_t(\tilde{U}_{x_i}) \leqq \Phi_t(V)$ for $t \geqq t_1$, for each i, so $\Phi_t(U) \leqq \Phi_t(V)$ for $t \geqq t_1$.

To obtain the stronger conclusion of the lemma, note that for each $x \in K_1$ there exists $\varepsilon_x > 0$ and a neighborhood W_x of x such that $\Phi([0, \varepsilon_x) \times W_x) \subset U$. As $\{W_x\}_{x \in K_1}$ is an open cover of K_1, there exists $x_1, x_2, \cdots, x_m \in K_1$ such that $K_1 \subset \bigcup_{i=1}^m W_{x_i}$. Let $U' = \bigcup_{i=1}^m W_{x_i}$ and $\varepsilon = \min_{1 \leqq i \leqq m} \varepsilon_{x_i}$. If $x \in U'$ and $0 \leqq s < \varepsilon$ then $x \in W_{x_i}$ for some i so $\Phi_s(x) \in U$. Thus $\Phi([0, \varepsilon) \times U') \subset U$ so $\Phi_s(U') \subset U$, $0 \leqq s < \varepsilon$. It follows that $\Phi_{t+s}(U') \subset \Phi_t(U) \leqq \Phi_t(V)$ for $t \geqq t_1$, $0 \leqq s < \varepsilon$.

Our main result is based on the following two results, which were first stated and proved by Hirsch [7], [8] for strongly monotone semiflows.

PROPOSITION 2.2 (Nonordering of limit sets). *An omega limit set cannot contain two ordered points.*

The following result is stated by Matano in [12] without proof.

PROPOSITION 2.3 (Limit set dichotomy). *If $x < y$ then either*

(a) $\omega(x) < \omega(y)$, *or*

(b) $\omega(x) = \omega(y) \subset E$.

We do not give detailed proofs of either Proposition 2.2 or Proposition 2.3 since this would involve repeating arguments very similar to those given by Hirsch in his proofs of corresponding results in [8, Thms. 6.2, 6.8]. A proof of Proposition 2.3 is obtained by proving analogues of the various results in Hirsch [8, § 6] and then arguing precisely as in Hirsch's proof of his limit set dichotomy. We state below those results, parallel to ones in [8], which are required for the proofs of Propositions 2.2 and 2.3. We will prove two of them to illustrate how our assumptions replace those by Hirsch [8].

Proposition 2.2 follows quite simply from the next result.

PROPOSITION 2.4 (Convergence criterion for strongly order-preserving flows). *If $\Phi_T(x) > x$ for some $T > 0$, then $\Phi_t(x) \to p \in E$ as $t \to \infty$.*

Proof. As in Lemma 6.3 of [8], $\omega(x)$ is an orbit of period T. Since Φ is strongly order preserving, there exist neighborhoods U of x and V of $\Phi_t(x)$ and $t_0 \geqq 0$ such that $\Phi_{t_0}(U) \leqq \Phi_{t_0}(V)$. For sufficiently small $\varepsilon > 0$, $\Phi_{T+s}(x) \in V$ for $0 \leqq s \leqq \varepsilon$ and hence $\Phi_{t_0}(x) < \Phi_{T+t_0+s}(x)$ for such s. This implies again that $\omega(\Phi_{t_0}(x)) = \omega(x)$ is an orbit of period $T+s$, $0 \leqq s \leqq \varepsilon$. Let $p \in \omega(x)$. Then $\Phi_{T+s}(p) = p$ for $0 \leqq s \leqq \varepsilon$. Hence $p = \Phi_s \Phi_T p = \Phi_s p$ for $0 \leqq s \leqq \varepsilon$. Let $t = n\varepsilon + s$ with $n \in N$ and $0 \leqq s < \varepsilon$. Then $\Phi_t p = \Phi_s \Phi_\varepsilon^n p = \Phi_s p = p$. Since any $t \geqq 0$ can be represented in this way, p is an equilibrium. As $\omega(x)$ is a periodic orbit, $\omega(x) = \{p\}$.

Simple modifications of Joel Friedman's argument in Proposition 6.6 of [8] yield Proposition 2.5.

PROPOSITION 2.5 (Colimiting principle). *If $x < y$ and for some sequence $t_k \to \infty$,*

$$\Phi_{t_k}(x) \to p, \quad \Phi_{t_k}(y) \to p, \quad \text{then } p \in E.$$

The next result follows immediately from Propositions 2.2 and 2.5 exactly as in Hirsch [8, Lemma 6.7].

PROPOSITION 2.6 (Intersection principle). *If $x < y$, then $\omega(x) \cap \omega(y) \subset E$.*

In addition to Proposition 2.6, two additional results are required for the proof of the limit set dichotomy. One of these is the following.

PROPOSITION 2.7 (Absorption principle). *Suppose that $x, y \in X$ and that neither are convergent points. If $\omega(x)$ contains a point $u < \omega(y)$, then $\omega(x) < \omega(y)$. If $\omega(x)$ contains a point $u > \omega(y)$, then $\omega(x) > \omega(y)$.*

The proof of the Absorption principle is similar to Hirsch's although we require an application of Lemma 2.1.

The next result is also used to prove the limit set dichotomy.

PROPOSITION 2.8 (Limit set separation principle). *Let x and y be as in Proposition 2.7 and $x < y$. If there exists $t_k \to \infty$ such that $\Phi_{t_k}(x) \to a$, $\Phi_{t_k}(y) \to b$, and $a < b$ then $\omega(x) < \omega(y)$.*

The proof of Proposition 2.8 is identical to that given by Hirsch [8, Lemma 6.10]. In particular, the proof requires the following result.

LEMMA 2.9. *Given the hypotheses of Proposition 2.8 then $\mathcal{O}^+(a) < b$.*

We give the proof of this lemma because it requires some modifications of Hirsch's corresponding proof [8, Lemma 6.10]. It also serves as another illustration of how, technically, our assumptions replace those of Hirsch.

Proof. For $u \in \overline{\mathcal{O}^+(x)}$, $v \in \overline{\mathcal{O}^+(y)}$, $u \leqq v$ define

$$\mathcal{J}(u, v) = \sup\{r \geqq 0: \Phi_t(u) \leqq v, 0 \leqq t \leqq r\}.$$

We verify two important properties of $\mathcal{J}$.

(P_1) $\mathcal{J}(\Phi_t(u), \Phi_t(v))$ is monotone nondecreasing in t.

It suffices to establish $\mathcal{J}(\Phi_t(u), \Phi_t(v)) \geqq \mathcal{J}(u, v)$. For $s \leqq \mathcal{J}(u, v)$, $\Phi_s(u) \leqq v$, so $\Phi_s\Phi_t(u) \leqq \Phi_t\Phi_s(u) \leqq \Phi_t(v)$. Hence

$$\mathcal{J}(\Phi_t(u), \Phi_t(v)) \geqq s \quad \text{for all } s \leqq \mathcal{J}(u, v).$$

(P_2) If $u_k \leqq v_k$, $u_k \in \overline{\mathcal{O}^+(x)}$, $v_k \in \overline{\mathcal{O}^+(y)}$ and $u_k \to u$, $v_k \to v$ then $\limsup_{k\to\infty} \mathcal{J}(u_k, v_k) \leqq \mathcal{J}(u, v)$.

Suppose that $\mathcal{J}(u, v) < \infty$ and $\limsup_{k\to\infty} \mathcal{J}(u_k, v_k) - \varepsilon > \mathcal{J}(u, v)$ for some $\varepsilon > 0$. Then $\mathcal{J}(u, v) + \delta < \mathcal{J}(u_{k_i}, v_{k_i})$ for all large i, for each δ, $0 \leqq \delta < \varepsilon$, where $\lim_{i\to\infty} \mathcal{J}(u_{k_i}, v_{k_i}) = \limsup_{k\to\infty} \mathcal{J}(u_k, v_k)$. It follows that $\Phi_{\mathcal{J}(u,v)+\delta}(u_{k_i}) \leqq v_{k_i}$ for all large i, $0 \leqq \delta < \varepsilon$. Letting $i \to \infty$, we obtain $\Phi_{\mathcal{J}(u,v)+\delta}(u) \leqq v$ for $0 \leqq \delta < \varepsilon$, a contradiction to the definition of $\mathcal{J}(u, v)$. Thus (P_2) holds.

It follows from (P_1) that $\alpha = \lim_{t\to\infty} \mathcal{J}(\Phi_t(x), \Phi_t(y))$ exists in $[0, \infty]$. By (P_2), $\mathcal{J}(a, b) \geqq \alpha$. Suppose the conclusion of the lemma does not hold, i.e., suppose $\mathcal{J}(a, b) < \infty$. For $0 \leqq r \leqq \mathcal{J}(a, b)$, $\Phi_r(a) \leqq b$. Actually, $\Phi_r(a) < b$, $0 \leqq r \leqq \mathcal{J}(a, b)$ since otherwise $a, b \in \omega(x)$ by invariance of $\omega(x)$ and $a < b$ gives a contradiction to Proposition 2.2. Let $K = \{\Phi_r(a): 0 \leqq r \leqq \mathcal{J}(a, b)\}$ so K is compact and $K < b$. By Lemma 2.1, there exist t_1, $\varepsilon > 0$ and open sets U and V with $K \subset U$ and $b \in V$ such that $\Phi_{t_1+\delta}(U) \leqq \Phi_{t_1}(V)$, $0 \leqq \delta \leqq \varepsilon$. It follows that $\Phi_{t_k}(y) \in V$ for $k \geqq k_0$ for some integer k_0. We claim that $\Phi_r\Phi_{t_k}(x) \in U$ for $0 \leqq r \leqq \mathcal{J}(a, b)$ and all large k. If not, we find a subsequence t_k' of t_k and a sequence $r_k \in [0, \mathcal{J}(a, b)]$ such that $\Phi_{r_k}\Phi_{t_k'}(x) \notin U$, $k = 1, 2, \cdots$. We may assume $r_k \to r \in [0, \mathcal{J}(a, b)]$ by passing to a subsequence if necessary. Letting $k \to \infty$, we obtain that $\Phi_r a \notin U$. But this is a contradiction to our choice of U and K and our claim is established.

Hence there exists k_2 such that

$$\Phi_{t_1+\delta}\Phi_r\Phi_{t_k}(x) \leqq \Phi_{t_1}\Phi_{t_k}(y),$$

for $k \geqq k_2$, $0 \leqq r \leqq \mathcal{J}(a, b)$, $0 \leqq \delta \leqq \varepsilon$. Thus,

$$\Phi_{r+\delta}\Phi_{t+t_k}(x) \leqq \Phi_{t+t_k}(y),$$

for $t \geqq t_1$, $k \geqq k_2$, $0 \leqq r \leqq \mathcal{J}(a, b)$, $0 \leqq \delta \leqq \varepsilon$. It follows that $\mathcal{J}(\Phi_{t+t_k}(x), \Phi_{t+t_k}(y)) \geqq \mathcal{J}(a, b) + \varepsilon$, $k \geqq k_2$. Letting $k \to \infty$, we obtain $\alpha \geqq \mathcal{J}(a, b) + \varepsilon$. But $\mathcal{J}(a, b) \geqq \alpha$ and this provides a contradiction. Hence $\mathcal{J}(a, b) = \infty$ and the conclusion of the lemma follows.

The proof of the limit set dichotomy (Proposition 2.3) can now be constructed exactly like Hirsch's original proof.

3. Convergence, quasi convergence, and stability. In this section we state and prove our main results. Several of these results (Theorems 3.3 and 3.10, and Corollary 3.12) overlap similar results of Hirsch [8] and Matano [11], [12]. In these cases, our results are more general and the proofs, we believe, are simpler. Essentially all the results of this section are based on Proposition 3.1 and Corollary 3.2, which, in turn, follow from the nonordering of limit sets (Proposition 2.2) and the limit set dichotomy (Proposition 2.3).

The following definitions will be used throughout this section. We say that $x \in X$ can be approximated from below (above) in X provided there exists a sequence x_n in X satisfying $x_n < x_{n+1} < x$ ($x < x_{n+1} < x_n$) and $x_n \to x$. If $x, y \in X$ then $[x, y] = \{z \in X: x \leqq z \leqq y\}$ is an order interval in X. A subset Y of X is order convex if $[y_1, y_2] \subset Y$ whenever $y_1, y_2 \in Y$ and $y_1 < y_2$.

We require the following compactness hypothesis for our semiflow Φ on X for all our results:

(C) For each $x \in X$, $\mathcal{O}^+(x)$ has compact closure in X. In addition, for each compact subset K of X, $\bigcup_{x\in K} \omega(x)$ has compact closure in X.

We assume that (C) holds throughout this section without further mention. This assumption is a relatively weak compactness requirement. It is certainly satisfied if, for example, every bounded (or even just compact) set $B \subset X$ has a bounded orbit $\mathcal{O}^+(B) = \bigcup_{x\in B} \mathcal{O}^+(x)$ and Φ_t is asymptotically smooth (see [4]). In particular, if bounded sets have bounded orbits and Φ_t is conditionally completely continuous for $t \geqq t_0$, $t_0 > 0$, then (C) holds. Recall that Φ_t is (conditionally) completely continuous if $\Phi_t(B)$ has compact closure in X whenever B is a bounded subset of X (and $\Phi_t(B)$ is a bounded set).

PROPOSITION 3.1 (Sequential limit set trichotomy). *Let X be an ordered metric space, and let Φ_t be a strongly order-preserving semiflow on X. Let $x_0 \in X$ have the property that it can be approximated from below in X by a sequence $\tilde{x}_n$. Then there exists a subsequence x_n of $\tilde{x}_n$ such that $x_n < x_{n+1} < x_0$, $n \geqq 1$, with $x_n \to x_0$ satisfying one of the following.*

(a) *There exists $u_0 \in E$ such that*

$$\omega(x_n) < \omega(x_{n+1}) < u_0 = \omega(x_0), \qquad n \geqq 1,$$

and

$$\lim_{n\to\infty} \operatorname{dist}(\omega(x_n), u_0) = 0.$$

(b) *There exists $u_0 \in E$ such that*

$$\omega(x_n) = u_0 < \omega(x_0), \qquad n \geqq 1.$$

If $u \in E$ and $u < \omega(x_0)$ then $u \leqq u_0$.

(c) *$\omega(x_n) = \omega(x_0) \subset E$ for $n \geqq 1$.*

Recall that $\operatorname{dist}(D, x) = \inf_{y\in D} d(y, x)$ gives the distance of a point $x \in X$ from a subset D of X.

Proof of Proposition 3.1. Let $\tilde{x}_n$ be any sequence satisfying $\tilde{x}_n < \tilde{x}_{n+1} < x_0$, $n \geqq 1$ and $\tilde{x}_n \to x_0$. By the limit set dichotomy (Prop. 2.3) either there exists a positive integer N such that $\omega(\tilde{x}_n) = \omega(\tilde{x}_m)$ for all m, n larger than N or there is a subsequence $\tilde{x}_{n_i}$ such that $\omega(\tilde{x}_{n_i}) < \omega(\tilde{x}_{n_{i+1}})$ for all i. By passing to this subsequence or renumbering the sequence, we may assume that either $\omega(x_n) = \omega(x_m)$ for all n, m or $\omega(x_n) < \omega(x_{n+1})$ for $n \geqq 1$, where x_n is the appropriate subsequence of $\tilde{x}_n$.

Suppose that the latter is the case. Then $\omega(x_n) < \omega(x_0)$ for all n. For if $\omega(x_n) = \omega(x_0)$ for some $n = n_0$ then $\omega(x_n) = \omega(x_0)$ for $n \geqq n_0$, a contradiction to $\omega(x_n) < \omega(x_{n+1})$. Let $\Omega = \{y\colon y = \lim y_n, y_n \in \omega(x_n)\} \subset \overline{\bigcup_{x\in K} \omega(x)}$ where $K = \{x_n\}_n \cup \{x_0\}$. By hypothesis, $\bigcup_{x\in K} \omega(x)$ has compact closure in X so Ω is nonempty and compact. Suppose y and u belong to Ω so that $y_n \to y$, $u_n \to u$ where $y_n, u_n \in \omega(x_n)$. Since $y_n < u_{n+1}$ and $u_n < y_{n+1}$ holds for all n, we obtain $y \leqq u$ and $u \leqq y$ so $u = y$. Thus Ω is a singleton, $\Omega = \{u_0\}$. Furthermore, Ω is invariant since each $\omega(x_n)$ is invariant. Thus $u_0 \in E$. It follows immediately from the definition of Ω and the fact that $\bigcup_{n\geqq 1} \omega(x_n)$ has compact closure in X that $\lim_{n\to\infty} \operatorname{dist}(\omega(x_n), u_0) = 0$. Finally, $\omega(x_n) < \omega(x_0)$ for all n implies that $u_0 \leqq \omega(x_0)$. If $u_0 \in \omega(x_0)$ then $\omega(x_0) = u_0$ by the nonordering principle for limit sets (Proposition 2.2). This is just (a) of the proposition. If $u_0 < \omega(x_0)$ then choose a neighborhood W of $\omega(x_0)$ and $t_0 \geqq 0$ such that $u_0 \leqq \Phi_t(W)$ for $t \geqq t_0$ (Lemma 2.1).

Now there exists $t_1 > 0$ such that $\Phi_{t_1}(x_0) \in W$ and by continuity of Φ_{t_1}, there is an integer n such that $\Phi_{t_1}(x_n) \in W$. It follows that $u_0 \leqq \Phi_t(x_n)$ for $t \geqq t_0 + t_1$. But $u_0 \in E$, so necessarily, $\omega(x_n) \geqq u_0$. On the other hand, $\omega(x_n) < \omega(x_{n+1}) \leqq u_0$ holds for every n so $\omega(x_n) < u_0$ holds for every n. This contradiction shows that $\omega(x_0) = u_0$. Thus (a) holds if $\omega(x_n) < \omega(x_{n+1})$, $n \geqq 1$.

Suppose now that $\omega(x_n) = \omega(x_1)$, $n \geqq 1$. Since $x_n < x_0$, the limit set dichotomy implies that either $\omega(x_n) = \omega(x_1) < \omega(x_0)$ or $\omega(x_n) = \omega(x_0)$ for $n \geqq 1$. The latter is precisely case (c) of the proposition. Suppose $\omega(x_1) < \omega(x_0)$. Let $u_0 \in \omega(x_1) = \omega(x_n) \subset E$ so $u_0 < \omega(x_0)$. By Lemma 2.1 there exists an open set W containing $\omega(x_0)$ and $t_0 \geqq 0$ such that $u_0 \leqq \Phi_t(W)$ for $t \geqq t_0$. Now, arguing exactly as in the previous paragraph, we obtain $u_0 \leqq \Phi_t(x_n)$ for some n and all large t. This yields $\omega(x_n) \geqq u_0$ and since $u_0 \in \omega(x_n)$ it follows that $\omega(x_n) = u_0$ by the nonordering principle for limit sets. Thus $\omega(x_1) = \omega(x_n) = u_0$ as asserted in case (b). Finally, if $u \in E$ and $u < \omega(x_0)$, we may argue exactly as above with u_0, that $u_0 = \omega(x_n) \geqq u$ for all large n, establishing that $u_0 \geqq u$.

Clearly, an analogous result holds if x_0 can be approximated from above in X.

In our next result we summarize some stability information which follows from the proof of Proposition 3.1.

COROLLARY 3.2. *Assume the hypotheses of Proposition* 3.1 *hold. In cases* (a), (b), *and* (c) *of Proposition* 3.1 *we have, in addition, the following*:

(a) *For any n there exists a neighborhood U_n of x_0 and $t_n \geqq 0$ such that*

$$\Phi_t(x_n) \leqq \Phi_t(U_n) \quad \text{for } t \geqq t_n.$$

(b) (i) *There exists a neighborhood O of u_0, t_0, $t_1 \geqq 0$, and n such that*

$$\Phi_t(O) \leqq \Phi_{t+t_1}(x_n) \quad \text{for } t \geqq t_0.$$

(ii) *There is a neighborhood U of x_0 with the following property: for each $x \in U$, $x < x_0$, there exists a neighborhood $V = V_x$ of x in U, an integer $N = N_x$ and $T = T_x > 0$ such that*

$$u_0 \leqq \Phi_t(V) \leqq \Phi_t(x_N), \quad \text{for } t \geqq T.$$

(c) *There is a neighborhood U of x_0 with the following property: for each $x \in U$, $x < x_0$, there exists a neighborhood $V = V_x$ of x in U and $T = T_x > 0$ such that*

$$\Phi_t(x_1) \leqq \Phi_t(V) \leqq \Phi_t(x_0) \quad \text{for } t \geqq T.$$

Moreover,

$$d(\Phi_t(x_0), \Phi_t(x_1)) \to 0 \quad \text{as } t \to \infty.$$

Proof. For (a), fix n. Since $\omega(x_n) < \omega(x_0)$, Lemma 2.1 implies the existence of neighborhoods $W_1 \supset \omega(x_n)$ and $W_2 \supset \omega(x_0)$ and $t_0 \geqq 0$ such that $\Phi_t(W_1) \leqq \Phi_t(W_2)$ for $t \geqq t_0$. There exists $t_1 > 0$ such that $\Phi_{t_1}(x_n) \in W_1$ and $\Phi_{t_1}(x_0) \in W_2$. By continuity of Φ_{t_1}, there is a neighborhood U_n of x_0 such that $\Phi_{t_1}(U_n) \subset W_2$. Hence $\Phi_{t+t_1}(x_n) \leqq \Phi_{t+t_1}(U_n)$ for $t \geqq t_0$, so (a) holds with $t_n = t_0 + t_1$.

Now consider (b). Since $u_0 < \omega(x_0)$, by Lemma 2.1, there exists a neighborhood W of $\omega(x_0)$, a neighborhood O of u_0, and $t_0 \geqq 0$ such that $\Phi_t(O) \leqq \Phi_t(W)$ for $t \geqq t_0$. There exists $t_1 > 0$ such that $\Phi_t(x_0) \in W$ for $t \geqq t_1$. By continuity of Φ_{t_1}, there exists n such that $\Phi_{t_1}(x_n) \in W$. It follows that

$$\Phi_t(O) \leqq \Phi_{t+t_1}(x_n), \qquad t \geqq t_0.$$

Choose a neighborhood U of x_0 such that $\Phi_{t_1}(U) \subset W$ and let $x \in U$ satisfy $x < x_0$. Then there exist neighborhoods V of x, $V \subset U$, $\mathcal{N}$ of x_0, and $t_2 \geqq 0$ such that $\Phi_t(V) \leqq \Phi_t(\mathcal{N})$ for $t \geqq t_2$. Choosing N such that $x_N \in \mathcal{N}$, we have $\Phi_t(V) \leqq \Phi_t(x_N)$ for $t \geqq t_2$. As $\Phi_{t_1}(V) \subset \Phi_{t_1}(U) \subset W$ we have from the paragraph above that

$$u_0 = \Phi(t)u_0 \in \Phi_t(O) \leqq \Phi_t(W) \leqq \Phi_t(\Phi_{t_1}(V)) = \Phi_{t+t_1}(V), \qquad t \geqq t_0,$$

hence

$$u_0 \leqq \Phi_t(V) \leqq \Phi_t(x_N), \qquad t \geqq t_0 + t_1 + t_2.$$

This establishes (b) with $T = t_0 + t_1 + t_2$.

Now consider (c). Since $x_1 < x_0$, there exists a neighborhood U of x_0 and $t_3 \geqq 0$ such that $\Phi_t(x_1) \leqq \Phi_t(U)$ for $t \geqq t_3$. Hence, if $x \in U$ and $x < x_0$ then there is a neighborhood V of x, $V \subset U$, and $t_4 \geqq 0$ such that $\Phi_t(V) \leqq \Phi_t(x_0)$ for $t \geqq t_4$. Since $V \subset U$, it follows that

$$\Phi_t(x_1) \leqq \Phi_t(V) \leqq \Phi_t(x_0) \quad \text{for } t \geqq t_3 + t_4.$$

If $d(\Phi_t(x_0), \Phi_t(x_1)) \to 0$ as $t \to \infty$ does not hold, then there exists $\varepsilon > 0$, $t_n \to \infty$ such that $d(\Phi_{t_n}(x_0), \Phi_{t_n}(x_1)) \geqq \varepsilon$ for $n \geqq 1$. Without loss of generality, we can assume $\Phi_{t_n}(x_0) \to u$, $\Phi_{t_n}(x_1) \to v$ where $u, v \in \omega(x_0)$. But $u \geqq v$ and nonordering of limit sets, Proposition 2.2, implies that $u = v$. This contradiction completes the proof of the corollary.

Remark 3.1. Corollary 3.2 contains stability information that will be exploited more fully later under the additional hypothesis that X is normally ordered (see Remark 3.2). A few remarks here will give the general idea.

(b) (i): Observe that if $x \in O$ and $x > u_0$ then $u_0 < \Phi_t(x) \leqq \Phi_{t+t_1}(x_n)$ for $t \geqq t_0$. Since $\Phi_t(x_n) \to u_0$ as $t \to \infty$, $\omega(x) = u_0$. Thus u_0 is "upper asymptotically stable" in the sense that $\omega(x) = u_0$ for all $x \in O$ with $x \geqq u_0$.

(b) (ii): Similarly, $\Phi_t(x_N) \to u_0$ as $t \to \infty$, so $\omega(v) = u_0$ for all $v \in V$. In particular, $\omega(x) = u_0$ for all $x \in U$ with $x < x_0$.

(c) Similar arguments imply that $\omega(x) = \omega(x_0)$ for all $x \in U$ with $x < x_0$.

THEOREM 3.3. *Let X be an ordered metric space, and let Φ_t be a strongly order-preserving semiflow on X. Suppose each point of X can be approximated from above or from below in X. Then $X = \operatorname{Int} Q \cup \overline{\operatorname{Int} C}$. In particular,* Int Q *is dense in X.*

Proof. Suppose $x_0 \in X \setminus \operatorname{Int} Q$. Then there exists a sequence $y_n \in X \setminus Q$ such that $y_n \to x_0$. By passing to a subsequence if necessary, we can assume that either y_n can be approximated from below in X for each n or y_n can be approximated from above in X for each n. We consider only the former case as the latter case is similar. Each y_n is the limit of a sequence $x_m^n \to y_n$, $x_m^n < x_{m+1}^n < y_n$. For each y_n, case (b) of Proposition 3.1 must hold since $y_n \notin Q$. By Corollary 3.2 and Remark 3.1, it follows that $y_n \in \overline{\operatorname{Int} C}$ for each n because $x_m^n \in \operatorname{Int} C$ for all large m. Hence $x_0 \in \overline{\operatorname{Int} C}$.

DEFINITION. If $x \in X$, then x is a stable point if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $d(\Phi_t(x), \Phi_t(y)) < \varepsilon$ for $t \geqq 0$ whenever $y \in X$ and $d(x, y) < \delta$. We let S be the subset of stable points of X. A point x is an asymptotically stable point if there is a neighborhood V of x with the property that for every $\varepsilon > 0$ there exists $t_\varepsilon > 0$ such that $d(\Phi_t(x), \Phi_t(y)) < \varepsilon$ if $t \geqq t_\varepsilon$ and $y \in V$. We let A denote the set of all asymptotically stable points.

Note that A is an open subset of X and that $A \subset S$. In fact, $A \subset S \subset Q$ under the hypotheses of Theorem 3.3. See also [8, Thm. 8.3 and § 9].

PROPOSITION 3.4. *If the hypotheses of Theorem* 3.3 *hold, then $S \subset Q$.*

Proof. If $x \in S$, then nearby points have nearby limit sets. It follows that only alternatives (a) and (c) of Proposition 3.1 are possible. Thus $x \in Q$.

DEFINITION. An ordered metric space X is called normally ordered if there exists a constant $k>0$ such that

$$d(u, v) \leqq k d(x, y)$$

for all x, y, u, v with $u, v \in [x, y]$.

Remark 3.2. If X is normally ordered, then in (b) and (c) of Corollary 3.2, each $x \in U$ with $x < x_0$ belongs to A. Indeed, the neighborhood V of x, in (b)(ii) of the corollary, can be taken for the neighborhood V in the definition of an asymptotically stable point. For if $y \in V$ then

$$d(\Phi_t(x), \Phi_t(y)) \leqq d(\Phi_t(x), u_0) + d(\Phi_t(y), u_0) \leqq 2kd(\Phi_t(x_N), u_0), \qquad t \geqq T.$$

Similarly, in case (c), if $y \in V$ then

$$\begin{aligned} d(\Phi_t(x), \Phi_t(y)) &\leqq d(\Phi_t(x), \Phi_t(x_0)) + d(\Phi_t(x_0), \Phi_t(y)) \\ &\leqq 2kd(\Phi_t(x_0), \Phi_t(x_1)), \qquad t \geqq T. \end{aligned}$$

Since the term on the right-hand side of each inequality tends to zero, it follows that $x \in A$. Similarly, the equilibrium u_0 in (b)(i) is upper asymptotically stable (see Hirsch [8]) in the sense that for every $\varepsilon > 0$ there exists $t_\varepsilon > 0$ such that if $x \in O$ and $x > u_0$ then $d(\Phi_t(x), u_0) < \varepsilon$ for $t > t_\varepsilon$.

THEOREM 3.5. *Let X be a normally ordered metric space, and let Φ_t be a strongly order-preserving semiflow on X. Assume that each $x \in X$ can be approximated from above or from below in X. Then $A \cup \operatorname{Int} C$ is dense in X.*

Proof. If $A \cup \operatorname{Int} C$ is not dense in X there exists an open set U in X such that $U \cap A = \varnothing = U \cap \operatorname{Int} C$. Let $x \in U$ and assume that x can be approximated from below. Then there is a sequence x_n such that $x_n < x_{n+1} < x$, $x_n \to x$ and one of the alternatives (a), (b), or (c) of Proposition 3.1 and Corollary 3.2 holds. We can assume $x_n \in U$ for all n. As $U \cap A = \varnothing$, only case (a) can hold (see Remark 3.2). Hence x is convergent. Since $x \in U$ was arbitrary, $U \subset C$. So $U \subset \operatorname{Int} C$ in contradiction to our assumption.

PROPOSITION 3.6. *Let X be an ordered metric space, and let Φ_t be a strongly order-preserving semiflow on X. Let $x_0 \in X$ be such that it can be approximated from above in X and from below in X. Then there exist sequences x_n and z_n in X satisfying $x_n \to x_0$, $z_n \to x_0$, $x_n < x_{n+1} < x_0 < z_{n+1} < z_n$, $n \geqq 1$, and one of the following holds*:

(a) *There exists $u_0 \in E$ such that, for $n \geqq 1$,*

$$\omega(x_n) < \omega(x_{n+1}) < \omega(x_0) = u_0 < \omega(z_{n+1}) < \omega(z_n)$$

and

$$\lim_{n\to\infty} \operatorname{dist}(\omega(x_n), u_0) = \lim_{n\to\infty} \operatorname{dist}(\omega(z_n), u_0) = 0.$$

(b) *There exists $u_0, v_0 \in E$ such that, for $n \geqq 1$, either*

(i) $\omega(x_n) < \omega(x_{n+1}) < \omega(x_0) = u_0 < v_0 = \omega(z_n)$,

$$\lim_{n\to\infty} \operatorname{dist}(\omega(x_n), u_0) = 0$$

and whenever $v \in E$, $v > u_0$ then $v \geqq v_0$, or

(ii) $\omega(x_n) = u_0 < v_0 = \omega(x_0) < \omega(z_{n+1}) < \omega(z_n)$

$$\lim_{n\to\infty} \operatorname{dist}(\omega(z_n), v_0) = 0$$

and whenever $u \in E$ and $u < v_0$ then $u \leqq u_0$.

(c) *There exists* $u_0 \in E$ *such that, for* $n \geqq 1$, *either*

(i) $\omega(x_n) < \omega(x_{n+1}) < \omega(x_0) = u_0 = \omega(z_n)$,

$$\lim_{n\to\infty} \operatorname{dist}(\omega(x_n), u_0) = 0$$

or

(ii) $\omega(x_n) = u_0 = \omega(x_0) < \omega(z_{n+1}) < \omega(z_n)$ *and*

$$\lim_{n\to\infty} \operatorname{dist}(\omega(z_n), u_0) = 0.$$

(d) *There exist equilibria* u_0 *and* v_0 *such that, for* $n \geqq 1$,

$$\omega(x_n) = u_0 < \omega(x_0) < v_0 = \omega(z_n).$$

If $u \in E$ *and* $u < \omega(x_0)$ *then* $u \leqq u_0$. *If* $v \in E$ *and* $\omega(x_0) < v$ *then* $v \geqq v_0$.

(e) *There exists* $u_0 \in E$ *such that, for* $n \geqq 1$, *either*

(i) $\omega(x_n) = u_0 < \omega(x_0) = \omega(z_n)$ *and, whenever* $u \in E$ *satisfies* $u < \omega(x_0)$, *then* $u \leqq u_0$, *or*

(ii) $\omega(x_n) = \omega(x_0) < u_0 = \omega(z_n)$ *and, whenever* $u \in E$ *satisfies* $u > \omega(x_0)$, *then* $u \geqq u_0$.

(f) *For* $n \geqq 1$, $\omega(x_n) = \omega(x_0) = \omega(z_n) \subseteq E$.

Proof. The proof is immediate from Proposition 3.1.

As in Proposition 3.1, the sequence $x_n(z_n)$ can be chosen to be a subsequence of any sequence $\tilde{x}_n(\tilde{z}_n)$ approximating x_0 from below (above) in X.

PROPOSITION 3.7. *Let the hypotheses of Proposition* 3.6 *hold. If* $x_0 \notin Q$, *then* (d) *of Proposition* 3.6 *holds. Moreover,* u_0 *and* v_0 *are upper and lower asymptotically stable equilibria, respectively. There is a neighborhood* U *of* x_0 *with the property that for each* $x \in U$ *with* $x < x_0$ $(x > x_0)$ *there exists a neighborhood* V *of* x *in* U *such that* $\omega(y) = u_0$ $(\omega(y) = v_0)$ *for every* $y \in V$. *If* $x \in U$, *then* $\omega(x) \subset [u_0, v_0]$. *If* X *is normally ordered, then* $x \in A$ *for each* $x \in U$ *satisfying* $x < x_0$ *or* $x > x_0$. *Finally, in addition to the above, assume that there exists an ordered Banach space* Y *with normal positive cone* Y_+ *such that* $X \subseteq Y$ *and the metric* d *and order* $\leqq$ *on* X *are inherited from the norm and partial order on* Y. *Assume that the order interval* $[u_0, v_0]$ *in* Y *is contained in* X *and that* Φ_t *is condensing on* $[u_0, v_0]$ *with respect to some measure of noncompactness for each* $t > 0$. *Then there exists* $w_0 \in E$ *satisfying* $u_0 < w_0 < v_0$.

Recall that Φ_t is condensing on $[u_0, v_0]$ if, for a suitable measure of noncompactness β, $\beta(\Phi_t(K)) < \beta(K)$ for any subset K of $[u_0, v_0]$, which is not relatively compact. See [21, § 11.3], for example.

Proof. It follows immediately that if $x_0 \notin Q$ then (d) of Proposition 3.6 holds since all other cases lead to $x_0 \in Q$. That $u_0(v_0)$ is upper (lower) asymptotically stable follows from Corollary 3.2 and Remark 3.2. Similarly, the existence of a neighborhood U of x_0 with the stated properties, follows from Corollary 3.2. If X is normally ordered, then the assertion that $x \in A$ if $x \in U$ and $x < x_0$ or $x > x_0$ follows from Remark 3.2.

It remains only to establish the final assertion. Suppose that the additional hypotheses contained in the last sentence of the proposition hold. Fix $t_0 > 0$ small and consider the map Φ_{t_0} restricted to $[u_0, v_0]$. We modify the proof of a result of Amann [1, Thm. 14.2, see also remarks on p. 667] using the fixed-point index on a suitable compact convex subset K of $[u_0, v_0]$ to establish the existence of a fixed point for Φ_{t_0} in $[u_0, v_0]$ different from u_0 and v_0.

First we show that there exists a compact convex subset K of $[u_0, v_0]$ containing both u_0 and v_0 and $\Phi_{t_0}(K) \subseteq K$. Following the proof of Sadovskii's Fixed-Point Theorem

in [21, § 11.5], we let $\mathscr{X}$ be the family of all closed convex subsets K of $[u_0, v_0]$ such that $u_0, v_0 \in K$ and $\Phi_{t_0}(K) \subseteq K$. Set

$$K = \cap\, \mathscr{X}, \qquad K_1 = \overline{\text{co}}\, \Phi_{t_0}(K).$$

Apparently, $K \in \mathscr{X}$ and $K_1 \subseteq K$. This implies $\Phi_{t_0}(K_1) \subseteq K_1$ so that $K_1 \in \mathscr{X}$. Hence $K = K_1 = \overline{\text{co}}\, \Phi_{t_0}(K)$. As Φ_{t_0} is condensing, the measure of noncompactness of K is zero.

On the convex compact set K the fixed-point index $i(g, U) = i_K(g, U)$ with the usual properties can be defined for continuous mappings $g: K \to K$ and for open (in the relative topology) subsets U of K. See [1].

By Corollary 3.2(b)(i) there exists $r > 0$ such that each point x of the closure of $B(u_0) = \{x \in K;\ d(u_0, x) < r\}$, respectively, $B(v_0) = \{x \in K;\ d(v_0, x) < r\}$, relative to K, satisfies $\omega(x) = u_0$, respectively, $\omega(x) = v_0$. Hereafter, all topological notions are to be understood relative to K. We will establish

$$i(\Phi_{t_0}, B(u_0)) = i(\Phi_{t_0}, B(v_0)) = i(\Phi_{t_0}, K) = +1,$$

which implies, by the additivity property of the fixed-point index, that Φ_{t_0} has a fixed point w_0 in $K \setminus (B(u_0) \cup B(v_0))$. Define the homotopy $F: [0, 1] \times \overline{B(u_0)} \to K$ by $F(\lambda, x) = \lambda u_0 + (1-\lambda)\Phi_{t_0}(x)$. If $F(\lambda, x) = x$ then $x - \Phi_{t_0}(x) = \lambda(u_0 - \Phi_{t_0}(x)) \leqq 0$ because $[u_0, v_0]$ is invariant under Φ. Equality can only hold if $x = u_0$ because u_0 is the only fixed point of Φ_{t_0} in $\overline{B(u_0)}$. Inequality implies $\Phi_{t_0}(x) > x$ that, by Proposition 2.4, gives that $\Phi_t(x)$ converges to an equilibrium larger than x. But this is impossible for any $x \in \overline{B(u_0)}$ because $\omega(x) = \{u_0\}$. We have shown that the fixed-point set of $F(\lambda, \cdot)$ in $\overline{B(u_0)}$ is precisely $\{u_0\}$. The homotopy property of the fixed-point index implies that

$$i(\Phi_{t_0}, B(u_0)) = i(F(1, \cdot), B(u_0)) = +1.$$

For the latter see [1]. Similar arguments yield the other equalities above. Thus there exists $w_0 \in K \setminus (B(u_0) \cup B(v_0))$ such that $\Phi_{t_0}(w_0) = w_0$. As $t_0 > 0$ was arbitrary, we obtain fixed points $w_n \in [u_0, v_0] \setminus (B(u_0) \cap B(v_0))$ of $\Phi_{2^{-n}}$ for all large n. The set $\{w_n\}$ is precompact in $[u_0, v_0]$ since each w_n is a fixed point of $\Phi_{2^{-j}}$ for all $n \geqq j$ and $\Phi_{2^{-j}}$ is condensing. A standard argument gives that a limit point of $\{w_n\}$ is an equilibrium of Φ in $[u_0, v_0] \setminus (B(u_0) \cup B(v_0))$: Let $w_{n_l} \to w$ for $l \to \infty$ and $t_l = 2^{-n_l}$, $\bar{w}_l = w_{n_l} \to w$. If $t > 0$, represent $t = m_l t_l + r_l$, $0 \leqq r_l < t_l$, with a nonnegative integer m_l, $l = 1, 2, \cdots$. Then

$$\Phi_t(w) = \lim_l \Phi_t(\bar{w}_l) = \lim_l \Phi_{r_l} \Phi_{m_l t_l}(\bar{w}_l) = \lim_l \Phi_{r_l}(\bar{w}_l) = w.$$

PROPOSITION 3.8. *Assume that the hypotheses of Proposition* 3.6 *hold and that X is normally ordered. If* (a) *of Proposition* 3.6 *holds then $x_0 \in S$. If* (*f*) *of Proposition* 3.6 *holds then $x_0 \in A$ and there exists a neighborhood U of x_0 such that $\omega(x) = \omega(x_0)$ for all $x \in U$.*

Proof. Suppose (a) of Proposition 3.6 holds. Since $\omega(x_n) < \omega(x_0) < \omega(z_n)$, Lemma 2.1 implies the existence of neighborhoods $W_1 \supset \omega(x_n)$ and $W_0' \supset \omega(x_0)$ and $t_n \geqq 0$ such that $\Phi_t(W_1) \leqq \Phi_t(W_0')$ for $t \geqq t_n$ and neighborhoods $W_0^2 \supset \omega(x_0)$ and $W_2 \supset \omega(z_n)$ and $s_n \geqq 0$ such that $\Phi_t(W_0^2) \leqq \Phi_t(W_2)$ for $t \geqq s_n$. If we let $U_n = W_0' \cap W_0^2$ and $T_n = \max\{t_n, s_n\}$ then $\Phi_t(W_1) \leqq \Phi_t(U_n) \leqq \Phi_t(W_2)$ for $t \geqq T_n$. Now we can argue exactly as in the proof of Corollary 3.2(a) that there exists $r_n \geqq T_n$ such that

$$\Phi_t(x_n) \leqq \Phi_t(U_n) \leqq \Phi_t(z_n) \quad \text{for } t \geqq r_n.$$

Since X is normally ordered it follows that

$$d(\Phi_t(x), \Phi_t(x_0)) \leqq k d(\Phi_t(x_n), \Phi_t(z_n)),$$

for $t \geqq r_n$, $x \in U_n$. Let $\varepsilon > 0$ be given and choose n such that $\operatorname{dist}(\omega(x_n), \omega(z_n)) < \varepsilon$. As $\Phi_t(x_n) \to \omega(x_n)$ and $\Phi_t(z_n) \to \omega(z_n)$ as $t \to \infty$, we can find $p_n > 0$ such that $d(\Phi_t(x_n), \Phi_t(z_n)) < 2\varepsilon$ if $t \geqq p_n$. Hence

$$d(\Phi_t(x), \Phi_t(x_0)) \leqq 2k\varepsilon,$$

for $t \geqq \max\{r_n, p_n\}$ if $x \in U_n$. By continuity of the semiflow, we may choose a neighborhood V_n of x_0, $V_n \subset U_n$, such that

$$d(\Phi_t(x), \Phi_t(x_0)) \leqq 2k\varepsilon, \qquad x \in V_n$$

for all $t \geqq 0$. It follows that $x_0 \in S$.

Suppose now that (f) of Proposition 3.6 holds. Arguing as in the proof of Corollary 3.2(c) we find a neighborhood U of x_0 and $T > 0$ such that

$$\Phi_t(x_1) \leqq \Phi_t(U) \leqq \Phi_t(z_1) \quad \text{for } t \geqq T.$$

Since X is normally ordered, we may conclude that

$$d(\Phi_t(x), \Phi_t(x_0)) \leqq k[d(\Phi_t(x_0), \Phi_t(x_1)) + d(\Phi_t(z_1), \Phi_t(x_0))],$$

for $x \in U$ and $t \geqq T$. Both $d(\Phi_t(x_0), \Phi_t(x_1)) \to 0$ and $d(\Phi_t(z_1), \Phi_t(x_0)) \to 0$ as $t \to \infty$ implying that $x \in A$.

THEOREM 3.9. *Let X be a normally ordered metric space and Φ_t be a strongly order-preserving semiflow on X. Suppose that each point x in X can be approximated from above or from below in X, and X contains an open and dense subset X_0 of points that can be approximated both from above and from below in X. Then $A \cup \operatorname{Int}(S \cap C)$ is dense in X.*

Proof. Recall the proof of Theorem 3.5. If U is open in X such that $A \cap U = \varnothing$, then $U \subset C$. But $\varnothing \neq U_0 := U \cap X_0 \subseteq S$. Actually, if $x \in U_0$, then $x \notin A$, so only alternative (a) of Proposition 3.6 can hold for x. Thus, by Proposition 3.8, $x \in S$. Hence $U_0 \subseteq \operatorname{Int}(S \cap C)$. These considerations imply that $U \cap [A \cup \operatorname{Int}(S \cap C)] \neq \varnothing$ for every open set U in X.

THEOREM 3.10 (Global Asymptotic Stability). *Let the hypotheses of Proposition 3.6 hold and suppose that every point of X can be approximated from above and from below in X. If X is connected and does not contain two ordered equilibria then X contains a unique equilibrium point which is the positive limit set of every orbit.*

Proof. Since no two points of E can be ordered, then for each $x_0 \in X$, only alternatives (a), (c), and (f) of Proposition 3.6 may hold. In particular, $x_0 \in Q$ for every $x_0 \in X$. But then no two limit sets $\omega(x_0)$, $\omega(x_1)$ can be ordered, $\omega(x_0) < \omega(x_1)$ or $\omega(x_1) < \omega(x_0)$ since each consists of equilibria. It follows that alternatives (a) and (c) of Proposition 3.6 cannot occur; only alternative (f) may occur. Fix $x_0 \in X$ and let $M = \{x \in X : \omega(x) = \omega(x_0)\}$. By Proposition 3.8, M is open in X. Similarly, the complement of M is open in X. Since X is connected and M is nonempty, it follows that $M = X$. Thus $\omega(x) = \omega(x_0) \subset E$ for all $x \in X$. If $u \in \omega(x_0)$ then $u = \omega(u) = \omega(x_0)$, so $\omega(x_0)$ is a single equilibrium. The proof is complete.

The hypothesis that no two equilibria are ordered is, of course, weaker than assuming that X contains exactly one equilibria, as assumed by Hirsch in a corresponding result [7], [8].

THEOREM 3.11. *Suppose that X is an order convex subset of a normally ordered Banach space Y. Let Φ_t be a strongly order-preserving semiflow on X such that Φ_t is condensing on every order interval in X for $t > 0$. Assume that any point in $X \setminus E$ can be approximated both from above and below and that X contains at most two order-related equilibria. Then $X = Q$.*

Proof. The proof is immediate from Proposition 3.7.

Remark 3.3. If X contains at most two equilibria, then we obtain from Theorem 3.11 that all points in X are convergent points. A typical application is $X = Y_+$ with zero being an equilibrium and X containing only one equilibrium different from zero. Then Theorem 3.11 implies that all trajectories converge to zero or to the other equilibrium. Actually a stronger alternative holds for trajectories starting in between the two equilibria.

COROLLARY 3.12. *Let $u < v$ be two elements in a normally ordered Banach space X, and let Φ_t be a strongly order-preserving condensing semiflow on the order interval $[u, v]$. Let u, v be the only equilibria of Φ_t on $[u, v]$. Then the following alternative holds. Either the trajectories starting in $[u, v]\backslash\{u, v\}$ all converge to u or they all converge to v.*

A similar result has been proved by Hirsch [8, Thm. 10.5]. Hirsch allows for topological vector spaces, but his compactness and order assumptions are stronger than ours and exclude examples like functional differential equations.

Proof of Corollary 3.12. Assume that there are elements $x, y \in [u, v]\backslash\{u, v\}$ with $\Phi_t(x) \to u$, $\Phi_t(y) \to v$, $t \to \infty$. As Φ is strongly order preserving, we find neighborhoods U, V of u, v and $t_0 > 0$ such that $\Phi_t(U) \leqq \Phi_t(x)$, $\Phi_t(V) \geqq \Phi_t(y)$ for $t \geqq t_0$. As X is normally ordered, we have that

$$\Phi_t(U \cap [u, v]) \to u, \quad \Phi_t(V \cap [u, v]) \to v, \quad t \to \infty.$$

The same proof as in Proposition 3.7 now implies that there is a third equilibrium between u and v in contradiction to our assumption.

Our next result provides sufficient conditions for the set of asymptotically stable points A to be dense in X. For this result, we need to assume that each point $x \in X$ belongs to some totally ordered arc in X. A subset J of X is a totally ordered arc provided $J = \psi(I)$ where I is a nontrivial interval in $\mathfrak{R}$, ψ is a continuous function from I into X, and $\psi(s) < \psi(t)$ whenever $s, t \in I$ and $s < t$.

THEOREM 3.13. *Suppose that X is a normally ordered metric space and that each $x \in X$ belongs to some totally ordered arc in X. Let Φ_t be a strongly order-preserving semiflow on X. If there does not exist a totally ordered arc of equilibria in X then A is dense in X.*

Proof. Assume that $x \in X\backslash\bar{A}$ and $x = \psi(0)$ with $\psi([0, 1])$ being a nontrivial totally ordered arc. In particular, $\psi([0, \varepsilon]) \subseteq X\backslash\bar{A}$ for some $\varepsilon > 0$. As in the proof of Theorem 3.9 ($U = X\backslash\bar{A}$), $\psi((0, \varepsilon))$ is a totally ordered arc of stable convergent points and $\omega(\psi(s)) \neq \omega(\psi(\tilde{s}))$ for $s \neq \tilde{s}$. Hence $\tilde{\psi}(s) = \omega(\psi(s)) = \{\tilde{\psi}(s)\}$ defines a continuous mapping from $(0, \varepsilon)$ into E forming a nontrivial totally ordered arc of equilibria in contradiction to our assumptions.

4. Ruling out nontrivial totally ordered connected sets of equilibria. To have the asymptotically stable points be dense we had to assume that there are no totally ordered arcs of equilibria. This assumption may be hard to verify for large systems of differential equations. A quite general condition is available, however, provided that the semiflow is analytic.

Let us make precise our scenario.

Scenario. We consider an open bounded set U in a Banach lattice X, with the cone X_+ having nonempty interior, and a monotone semiflow Φ on $\bar{U}$.

Let $X + iX$ be the complexification of X. We assume that, for any equilibrium $x_0 \in U$ there exist a neighborhood V of 0 in X and numbers $0 < \varepsilon < t_0$ such that Φ can be extended to a continuous mapping from $(t_0 - \varepsilon, t_0 + \varepsilon) \times (x_0 + V + iV)$ to $X + iX$.

We further assume that, for any $t \in (t_0-\varepsilon, t_0+\varepsilon)$, Φ_t is complex differentiable (i.e., analytic) as a mapping from x_0+V+iV to $X+iX$ and that

$$D\Phi_t(x) \in L(X+iX)$$

depends continuously on X uniformly in $t \in (t_0-\varepsilon, t_0+\varepsilon)$. Here $L(X+iX)$ denotes the complex Banach space of bounded linear operators on $L(X+iX)$ endowed with the uniform operator norm.

Finally, we assume that, for any $t>0$, Φ_t is continuously differentiable as a mapping from U to U and that $D\Phi_t: U \to L(X)$ can be continuously extended to $\bar{U}$. Furthermore, we assume that, for any equilibrium $x_0 \in \bar{U}$, $D\Phi_t(x_0)$, $t \geqq 0$, forms an irreducible strongly continuous semigroup on X such that the essential type $\omega_{\text{ess}}(x_0)$ (or essential growth bound) is strictly negative.

We recall that

$$\omega_{\text{ess}}(x_0) = \inf_{t>0} \frac{1}{t} \ln |D\Phi_t(x_0)|_\beta = \lim_{t\to\infty} \frac{1}{t} \ln |D\Phi_t(x_0)|_\beta,$$

with $|\cdot|_\beta$ denoting the measure of noncompactness. See § 8 of [3]. Furthermore,

$$r_{\text{ess}}(D\Phi_t(x_0)) = e^{w_{\text{ess}}(x_0)t}, \qquad t \geqq 0,$$

with r_{ess} denoting the essential spectral radius. $\omega_{\text{ess}}(x_0) = -\infty$, e.g., if $D\Phi_t(x_0)$ is compact for some $t>0$. Before we formulate the main result of this section let us recall some facts about irreducible strongly continuous semigroups.

Let $A(x_0)$ be the infinitesimal generator of $D\Phi_t(x_0)$, with x_0 being an equilibrium. Then the *spectral bound* $s(x_0)$ of $A(x_0)$ is defined by

$$s(x_0) = \sup \{\text{Re } \lambda; \lambda \text{ spectral value of } A(x_0)\}.$$

It is well known that $s(x_0)$ is a spectral value itself if $s(x_0) > -\infty$. See, e.g., Theorem 8.7 of [3].

We are now able to formulate our main result.

THEOREM 4.1. *Consider the scenario outlined above. Assume further that any closed set of equilibria in $\bar{U}$ is compact and that*

$$s(x_0) \neq 0,$$

for any equilibrium $x_0 \in \partial U$ (if there are any). Then every totally ordered connected set of equilibria in $\bar{U}$ is a singleton (i.e., consists of one element only).

It is convenient to prove the following result first.

LEMMA 4.2. *Consider the scenario outlined at the beginning of this section. Let D be a totally ordered connected set of equilibria in $\bar{U}$ which consists of more than one element. Then $s(x_0) = 0$ for any $x_0 \in D$.*

Proof. Let $x_0 \in D$. Then x_0 can be approximated by other elements in D from above or from below. We assume that there are elements $x_n > x_0$ such that $x_n \to x_0$. Recall that x_0, x_n are equilibria. Thus, defining

$$u_n = \frac{1}{\|x_n - x_0\|}(x_n - x_0),$$

we obtain that

$$u_n - D\Phi_t(x_0)u_n \to 0 \quad \text{for } n \to \infty, \quad t \geqq 0.$$

Let β be the measure of noncompactness. See § A.3.1 of [3]. Then

$$\beta(\{u_n; n \in N\}) \leqq |D\Phi_t(x_0)|_\beta \beta(\{u_n\}_n) \leqq (M e^{(w_{\text{ess}}(x_0)+\varepsilon)t})\beta(\{u_n; n \in N\}),$$

with $\varepsilon > 0$ such that $w_{ess}(x_0) + \varepsilon < 0$ and M depending on ε. Hence we realize from choosing t large enough that

$$\beta(\{u_n;\, n \in \mathbf{N}\}) = 0.$$

This implies that the closure of $\{u_n;\, n \in \mathbf{N}\}$ is compact, and we find $v \geqq 0$, $\|v\| = 1$ such that

$$D\Phi_t(x_0)v = v \quad \text{for all } t \geqq 0.$$

Clearly, v is an eigenvector of $A(x_0)$ belonging to the eigenvalue zero. As $D\Phi_t(x_0)$, $t \geqq 0$, is irreducible, $v \in \operatorname{Int} X_+$. This implies that the spectral radius of $D\Phi_t(x_0) \leqq 1$, hence $s(A) \leqq 0$. See Proposition 8.6 and the preceding remarks of [3]. As zero is an eigenvalue of A, $s(A) = 0$.

Remark 4.1. It follows from our assumptions that $s(x_0) = 0$ is a pole of the resolvent of $A(x_0)$. The irreducibility of $D\Phi_t(x_0)$ implies that $s(x_0)$ is a simple eigenvalue of $A(x_0)$ with an eigenvector $v \in \operatorname{Int} X_+$ and a strictly positive eigenfunctional $v^* \in X^*$ of $A^*(x_0)$. If we normalize v^* and v such that $\|v\| = 1$, $v^*(v) = 1$, then

$$Px = v^*(x)v$$

is the projection onto the eigenspace of zero. In particular, there exists $\varepsilon > 0$, $M > 0$ such that

$$\|(I - P)D\Phi_t(x_0)x\| = \|D\Phi_t(x_0)x - v^*(x)v\| \leqq M e^{-\varepsilon t}\|x\|,$$

$t \geqq 0$. See Theorem 9.11 of [3]. Moreover, v is the only eigenvector of $A(x_0)$ in X_+.

We are now able to give the proof of Theorem 4.1.

Proof of Theorem 4.1. By Zorn's lemma we can assume that D is a maximal totally ordered connected set of equilibria in $\bar{U}$. Let D contain at least two points. By maximality we know that D is closed. Hence D is compact by assumption. $D \cap \partial U = \varnothing$, otherwise we obtain a contradiction from Lemma 4.2 and the assumption that $s(x_0) \neq 0$ for any $x_0 \in \partial U$. As D is compact and $\operatorname{Int} X_+ \neq \varnothing$, $x_0 = \sup D$ exists. See Schaefer [16, II, Prop. 7.6]. We intend to derive a contradiction to the maximality of D. As D is totally ordered, $x_0 \in D$. Let v, v^* be the eigenvectors of $A(x_0)$, $A^*(x_0)$ associated with the eigenvalue $s(x_0) = 0$, v, $v^* > 0$, $\|v\| = 1$, $v^*(v) = 1$ and

$$Px = v^*(x)v.$$

See Remark 4.1. Set $Q = I - P$. We intend to apply the uniform contraction principle in order to find unique fixed points $u = u(t, z) \in QX$ satisfying

$$u = Q\Phi_t(x_0 + zv + u) - Qx_0 =: F(t, z, u).$$

See Theorem 2.2 of [2, Chap. 1].

Note that

$$D_uF(t, z, u) = QD\Phi_t(x_0 + zv + u),$$

and

$$D_uF(t, 0, 0) = QD\Phi_t(x_0),$$

with $D_uF(t, z, u)$ being a complex derivative in $X + iX$ for $t \in (t_0 - \varepsilon, t_0 + \varepsilon)$, $zv + u \in V + iV$. See the corresponding assumption in the scenario. Note that $QD\Phi_t(x_0)$ is a strongly continuous semigroup on $Q(X + iX)$. Changing the norm in $Q(X + iX)$ equivalently we can assume that

$$\|D_uF(t, 0, 0)\| < e^{-\delta t}, \qquad t > 0,$$

with $\delta > 0$. See Remark 4.1 and Theorem 5.2 of [13, Chap. 1]. As we have assumed that $D\Phi_t(x)$ is a continuous function of x uniformly for $|t - t_0| < \varepsilon$, we can achieve that

$$\|D_u F(t, z, u)\| < 1$$

in a neighborhood of $(t_0, 0, 0)$.

Hence $u \mapsto F(t, z, u)$ is a contraction in a neighborhood $W + iW$ of $0 + i0$, uniformly for $|t - t_0| < \varepsilon$, z in a neighborhood V_1 of 0, $V_1 \subseteq \mathbf{C}$. Note that, by our assumptions, $F(t, \cdot, \cdot)$ is analytic in $V_1 \times (W + iW)$. Using the metric structure only, the uniform contraction principle provides us with a unique solution $u = u(t, z)$ of

$$u = F(t, z, u), \qquad u(t_0, 0) = 0,$$

if $|t - t_0| < \varepsilon$, $z \in V_1$, $u \in W + iW$. After possibly having chosen ε and V_1 somewhat smaller than before, $u(t, z)$ continuously depends on (t, z). Proceeding as in the proof of Theorem 2.2 in [2, Chap. 1], we find that, for $|t - t_0| < \varepsilon$, $u(t, \cdot)$ is analytic in $V_1 \subseteq C$. Furthermore,

$$\partial_z u(t, z) = (I_{Q(X+iX)} - D_u F(t, z, u(t, z)))^{-1} QD\Phi_t(x_0 + zv + u(t, z))v.$$

Hence, as $u(t_0, 0) = 0$ and $QD\Phi_t(x_0)v = Qv = 0$, we have

$$\partial_z u(t_0, 0) = 0.$$

Returning to our connected closed set D of equilibria, $x_0 = \sup D$, we find that

$$x = x_0 + v^*(x - x_0)v + Q(x - x_0),$$

for $x \in D$, and

$$Q(x - x_0) = F(t, v^*(x - x_0), Q(x - x_0)).$$

Uniqueness implies that

$$Q(x - x_0) = u(t, v^*(x - x_0)),$$

if $|t - t_0| < \varepsilon$, $x \in D$, $\|x - x_0\| < \varepsilon$, if $\varepsilon > 0$ is chosen small enough. Hence

$$u(t, \sigma) - u(t_0, \sigma) = 0,$$

for $|t - t_0| < \varepsilon$, $-\varepsilon < \sigma \leqq 0$ if $\varepsilon > 0$ is chosen small enough. Recall that D is totally ordered, connected, and contains more than just one point. As $u(t, \sigma) - u(t_0, \sigma)$ is an analytic function of σ, we obtain that

$$u(t, \sigma) = u(t_0, \sigma),$$

for $|t - t_0| < \varepsilon$, $|\sigma| < \varepsilon$, if $\varepsilon > 0$ is chosen small enough. Hence, by the definition of F,

$$u(t_0, \sigma) + Qx_0 = Q\Phi_t(x_0 + \sigma v + u(t_0, \sigma)),$$

for $|t - t_0| < \varepsilon$, $|\sigma| < \varepsilon$.

We claim that $x(\sigma) = x_0 + \sigma v + u(t_0, \sigma)$ is a fixed point of Φ_t if $|t - t_0| < \varepsilon$. To this end we still have to show that

$$v^*\Phi_t(x_0 + \sigma v + u(t_0, \sigma)) - v^*(x_0) - \sigma = 0,$$

for $|t - t_0| < \varepsilon, |\sigma| < \varepsilon$. But this actually holds because it is true for $\sigma < 0, |\sigma| < \varepsilon, |t - t_0| < \varepsilon$ and the left-hand side of this equation is analytic in σ. Hence $x(\sigma)$ is a fixed point of Φ_t, if $|\sigma| < \varepsilon$, $|t - t_0| < \varepsilon$. Now

$$x(\sigma) = \Phi_{t_0+r}(x(\sigma)) = \Phi_r \Phi_{t_0}(x(\sigma)) = \Phi_r(x(\sigma)),$$

for $0 \leqq r < \varepsilon$. Applying the semigroup property again we find that

$$\Phi_t(x(\sigma)) = x(\sigma), \qquad t \geqq 0.$$

As $\partial_z u(t_0, 0) = 0$ and $v \in \operatorname{Int} X_+$, we finally note that

$$x(\sigma) = x_0 + \sigma v + u(t_0, \sigma)$$

is a strictly increasing function of $|\sigma| < \varepsilon$, if $\varepsilon > 0$ is chosen small enough. Hence the totally ordered continuum of equilibria D can be properly extended, in contradiction to its maximality. This proves our theorem.

REFERENCES

[1] H. AMANN, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620-709.

[2] S.-N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Grundlehren Math. Wiss. 251, Springer-Verlag, Berlin, New York, 1982.

[3] PH. CLÉMENT, H. J. A. M. HEIJMANS, S. ANGENENT, C. J. VAN DUIJN, AND B. DE PAGTER, *One-Parameter Semigroups*, CWI Monograph 5, North-Holland, Amsterdam, 1987.

[4] J. K. HALE, *Asymptotic Behavior of Dissipative Systems*, Math. Surveys Monographs, 25, American Mathematical Society, Providence, RI, 1988.

[5] M. W. HIRSCH, *Systems of differential equations which are competitive or cooperative* I: *limit sets*, SIAM J. Math. Anal., 13 (1982), pp. 167-179.

[6] ———, *Systems of differential equations which are competitive or cooperative* II: *convergence almost everywhere*, SIAM J. Math. Anal., 16 (1985), pp. 432-439.

[7] ———, *The dynamical systems approach to differential equations*, Bull. Amer. Math. Soc., 11 (1984), pp. 1-64.

[8] ———, *Stability and convergence in strongly monotone dynamical systems*, J. Reine Angew. Math., 383 (1988), pp. 1-53.

[9] R. H. MARTIN, JR. AND H. L. SMITH, *Abstract functional differential equations and reaction-diffusion systems*, Trans. Amer. Math. Soc., to appear.

[10] ———, *Reaction diffusion systems with time delays: monotonicity, invariance, comparison and convergence*, preprint.

[11] H. MATANO, *Strongly order preserving local semi-dynamical systems—theory and applications*, in Semigroups, Theory and Applications. Vol. I, H. Brezis, M. G. Crandall, and F. Kappel, eds., Research Notes in Mathematics 141, Pitman, Boston, pp. 178-185; Longman Scientific & Technical, London, 1986.

[12] ———, *Strong comparison principle in nonlinear parabolic equations*, in Nonlinear Parabolic Equations: Qualitative Properties of Solutions, L. Boccardo and A. Tesei, eds., Res. Notes in Math. 149, Pitman, Boston, pp. 148-155; Longman Scientific & Technical, London, 1987.

[13] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.

[14] P. POLÁČIK, *Convergence in smooth strongly monotone flows defined by semilinear parabolic equations*, preprint.

[15] ———, *Domains of attraction of equilibria and monotonicity properties of convergent trajectories in parabolic systems admitting strong comparison principles*, preprint.

[16] H. H. SCHAEFER, *Banach Lattices and Positive Operators*, Grundlehren Math. Wiss. 215, Springer-Verlag, Berlin, New York, 1974.

[17] J. F. SELGRADE, *A Hopf bifurcation in single-loop positive-feedback systems*, Quart. J. Appl. Math., 40 (1982), pp. 347-351.

[18] H. L. SMITH, *Systems of ordinary differential equations which generate an order preserving flow. A survey of results*, SIAM Rev., 30 (1988), pp. 87-111.

[19] ———, *Monotone semiflows generated by functional differential equations*, J. Differential Equations, 66 (1987), pp. 420-442.

[20] H. L. SMITH AND H. R. THIEME, *Monotone semiflows in scalar nonquasimonotone functional differential equations*, J. Math. Anal. Appl., to appear.

[21] E. ZEIDLER, *Nonlinear Functional Analysis and Its Applications* I: *Fixed Point Theorems*, Springer-Verlag, Berlin, New York, 1986.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 693-720, May 1990
© 1990 Society for Industrial and Applied Mathematics
008

HOMOCLINIC BIFURCATIONS WITH NONHYPERBOLIC EQUILIBRIA*

BO DENG†

Abstract. A general geometric approach is given for bifurcation problems with homoclinic orbits to nonhyperbolic equilibrium points of ordinary differential equations. It consists of a special normal form called admissible variables, exponential expansion, strong λ-lemma, and Lyapunov-Schmidt reduction for the Poincaré maps under Sil'nikov variables. The method is based on the Center Manifold Theory, the contraction mapping principle, and the Implicit Function Theorem.

Key words. homoclinic orbit, admissible variables, exponential expansion, strong λ-lemma, center manifold, saddle-node bifurcation

AMS(MOS) subject classifications. 34A34, 34C25, 34C28

1. Introduction. In this paper we will study homoclinic bifurcations with nonhyperbolic equilibrium points. The method we will introduce consists of four parts: a special normal form theory, exponential expansions for the Sil'nikov solution, the strong λ-lemma, and Lyapunov-Schmidt reduction for the Poincaré maps under Sil'nikov variables. Let us begin with a survey on the same method with hyperbolic equilibria. Hopefully, this will help us develop the right intuition to the problems we have in mind.

Consider a system of ordinary differential equations

$$\dot{u} = F(u), \qquad u \in \mathbb{R}^d, \tag{1.1}$$

where F is C^r and r is large enough so that whenever C^{r-k} appears, we have $r-k \geqq 1$. Suppose the origin $u=0$ is a hyperbolic equilibrium point. Let $1 \leqq m \leqq d$ and $1 \leqq n \leqq d$ with $m+n=d$ be the numbers of the eigenvalues with negative and positive real parts, respectively, for the linearization $DF(0)$. Then, up to a linear change of coordinates, we may assume $u=(x, y)$, $DF(0) = \operatorname{diag}(A, B)$, and

$$\dot{x} = Ax + f(x, y), \qquad \dot{y} = By + g(x, y), \tag{1.2}$$

which satisfy that the real parts of the eigenvalues for the $m \times m$ and $n \times n$ matrices A and B are negative and positive, respectively, and f, g are vanishing at the origin together with their first derivatives.

Given a triplet (τ, x_0, y_1), a solution $(x, y)(t)$ is a solution to the Sil'nikov problem if the conditions $x(0)=x_0$ and $y(\tau)=y_1$ are satisfied. Interpreted geometrically in Fig. 1.1, it shows that for a given initial coordinate surface $x=x_0$, an end coordinate surface $y=y_1$, and a time τ, a Sil'nikov solution takes exactly τ units of time to travel from $x=x_0$ to $y=y_1$. Observe that when $\tau=0$ this problem reduces to the initial value problem. Thus it is not surprising to expect that the Sil'nikov solution is existing, unique, and continuously differentiable in its Sil'nikov data τ, x_0, and y_1. To be more precise, let $B(\delta) \overset{\text{def}}{=} \{(x, y) \,|\, |x| \leqq \delta, |y| \leqq \delta\}$ be the box neighborhood of the origin; then there exists a small δ_0 such that for every triplet $(\tau, x_0, y) \in \mathbb{R}^+ \times B(\delta_0)$ there exists a unique Sil'nikov solution $(x, y)(t) \overset{\text{def}}{=} (x, y)(t, \tau, x_0, y_1)$ in $B(2\delta_0)$ for $0 \leqq t \leqq \tau$. This solution is Lipschitz in the Sil'nikov data τ, x_0, and y_1 if the nonlinear terms f and g are Lipschitz, or C^r if they are C^r. The proof easily follows from the uniform contraction

* Received by the editors December 12, 1988; accepted for publication (in revised form) May 15, 1989.
† Department of Mathematics and Statistics, University of Nebraska, Lincoln, Nebraska 68588-0323.

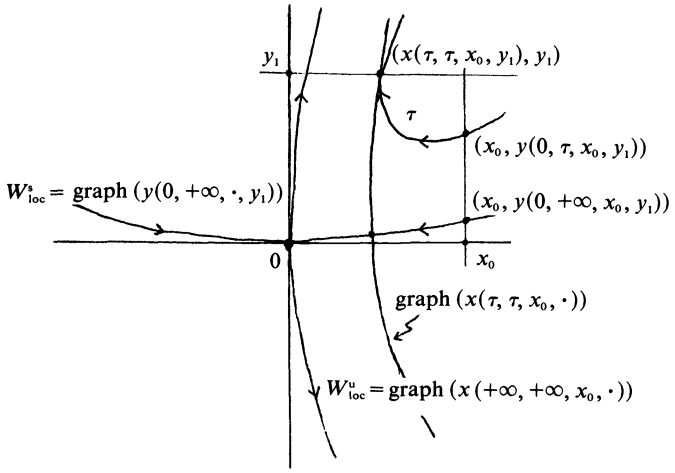


FIG. 1.1. *The hyperbolic structure in terms of the Sil'nikov solutions.*

mapping theorem together with the following equivalent integral equations:

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-s)}f(x(s), y(s))\, ds,$$
$$y(t) = e^{B(t-\tau)}y_1 + \int_\tau^t e^{B(t-s)}g(x(s), y(s))\, ds. \tag{1.3}$$

Also, the hyperbolicity is crucial for the validity of all $\tau \geqq 0$ for it implies the exponential functions are all bounded in the formula above. See Sil'nikov (1967) and Deng (1988a–d) for details. The first natural task is to formulate the Sil'nikov problem when the nonhyperbolicity is taken into consideration.

Before answering this question, let us see first how the Sil'nikov solution can give us a better understanding of the intrinsic local structure near a hyperbolic equilibrium. For instance, the stable and unstable manifolds can be described by the limiting behaviors of the functions $y(0, \tau, x_0, y_1)$ and $x(\tau, \tau, x_0, y_1)$ as $\tau \to +\infty$. Indeed, because the initial point for a given trajectory is given as $(x_0, y(0, \tau, x_0, y_1))$, the uniqueness of the stable manifold will imply that the family of functions $y(0, \tau, \cdot, \cdot)$ converges in the C^r-topology as $\tau \to +\infty$ and that the limit, denoted as $y(0, +\infty, \cdot, \cdot)$, does not depend on the y_1 variable. Moreover, the local stable manifold W^s_{loc} is precisely the graph of the C^r function $y(0, +\infty, \cdot, y_1)$ for any fixed y_1 since the trajectory through $(x_0, y(0, +\infty, x_0, y_1))$ stays in the box $B(2\delta_0)$ for all the positive time (cf. Fig. 1.1). Similarly, we have $W^u_{loc} = \text{graph}\,(x(+\infty, +\infty, x_0, \cdot))$, where $x(+\infty, +\infty, \cdot, \cdot)$ is the limit function of $x(\tau, \tau, \cdot, \cdot)$ as $\tau \to +\infty$. As another example, observe that the image of the n-dimensional "straight" disc $x = x_0$ under the time τ mapping of the flow is a curved manifold given as graph $(x(\tau, \tau, x_0, \cdot))$, which converges to the unstable manifold in the C^r-topology sense. In fact, this simple observation is just a special case of the so-called λ-lemma, or inclination lemma, for any n-dimensional disc transversely intersecting the stable manifold. This C^r λ-lemma can be proved by directly using the Implicit Function Theorem and the C^r-convergence of the functions $x(\tau, \tau, \cdot, \cdot)$ and $y(0, \tau, \cdot, \cdot)$ as $\tau \to +\infty$. For the complete details, see Deng (1988c).

Of the most importance is to incorporate the Sil'nikov solution into our studies of homoclinic bifurcations. To do this, let us assume that there is a homoclinic orbit

Γ to the origin and consider a Poincaré map Π around the orbit. Figure 1.2 heuristically illustrates the construction of such a Π. Here, Σ_0 and Σ_1 are two $(d-1)$-dimensional Poincaré cross sections in $B(2\delta_0)$ with the property that they are transverse to $\Gamma \cap W^s_{loc}$ and $\Gamma \cap W^u_{loc}$, respectively. For simplicity, let us assume $\Sigma_0 = \{x^{(1)} = \delta_0\}$ and $\Sigma_1 = \{y^{(1)} = \delta_0\}$ locally. σ_0 is the set of those points $p = (x_0, y_0)$ of Σ_0 whose local trajectories hit Σ_1 at $q = (x_1, y_1)$ at the first time $\tau = \tau(p)$. Thus the local map Π_0 is defined on σ_0 with the rule $p \to q \in \Sigma_1$. The global map Π_1 is defined in the same way by following the trajectories from Σ_1 back to Σ_0. Without loss of generality, however, Σ_1 can be taken as the domain for Π_1 and all trajectories starting from Σ_1 take roughly a constant time to reach Σ_0. In contrast, the domain σ_0 of the local map is a proper subset of Σ_0, not containing any point from the local stable manifold, and the time τ diverges to infinity as the initial point p approaches the stable manifold. The Poincaré map is now defined as $\Pi = \Pi_1 \circ \Pi_0$.

In general, the Poincaré map Π is difficult to deal with directly due to the long-time behavior of the local flow. Thus we wish to find a new variable for the Poincaré map such that it becomes tractable in terms of the new variable. The Sil'nikov data (τ, x_0, y_1) serves us precisely for this purpose. We will see this more clearly later on, but for the moment let us note that $\Delta = \{(\tau, x_0, y_1) \mid \tau \geqq 0, x_0^{(1)} = \delta_0, y_1^{(1)} = \delta_0, |x_0| \leqq \delta_0$, and $|y_1| \leqq \delta_0\}$ is imbedded in $\mathbb{R}^{d-1}$, and that the mapping $\rho_0 : \Delta \to \sigma_0$ with $(\tau, x_0, y_1) \to (x_0, y(0, \tau, x_0, y_1))$ gives rise to a C^r change of variables since its inverse can be easily defined by: $(x_0, y_0) \to (\tau, x_0, y_1)$, where $(x_0, y_0) = p \in \sigma_0$, $(x_1, y_1) = q = \Pi_0(p_0) \in \Sigma_1$ with $\tau = \tau(p)$. (τ, x_0, y_1) is called the Sil'nikov variable and the change of variables ρ_0 transforms these otherwise intractable variables τ, x_0, and y_1 into independent variables. Moreover, also note that the local map in the new variable is now simply given as $(x(\tau, \tau, x_0, y_1), y_1) = \Pi_0(\rho_0(\tau, x_0, y_1)) \overset{\text{def}}{=} \rho_1(\tau, x_0, y_1)$. Also, the fixed point of Π, for example, is now equivalent to solving the equation $\rho_0(\tau, x_0, y_1) = \Pi_1(\rho_1(\tau, x_0, y_1))$ for $(\tau, x_0, y_1) \in \Delta$.

However, the property of the uniform convergences of $x(\tau, \tau, x_0, y_1)$ and $y(0, \tau, x_0, y_1)$ as $\tau \to +\infty$ alone is not enough to make full use of the nice representations above for the local map of the flow. This is because the intersection of the stable and unstable manifolds along a homoclinic orbit must not be transverse. But, on the other

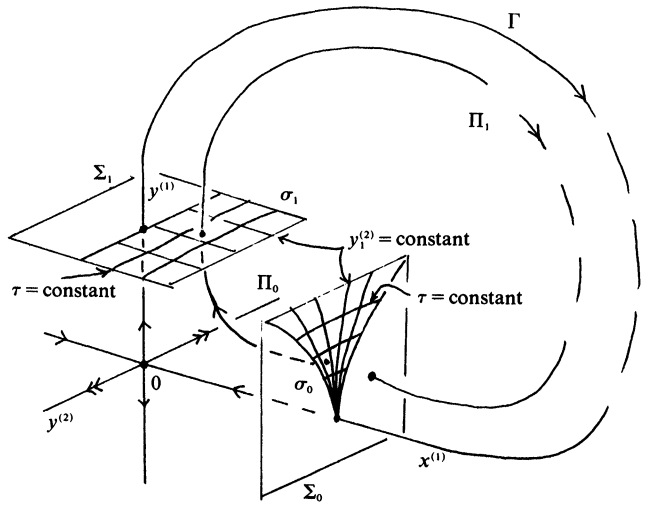


FIG. 1.2. *The Poincaré map for flows and the Sil'nikov variables.*

hand, it is quite sufficient for studying the dynamics of a transverse homoclinic point for diffeomorphisms. To see this, we refer our readers to Sil'nikov (1967) and Moser (1973). It turns out that to compensate for this loss of transversality in vector fields, we need to know a finer and subtler structure of the local flow on an exponentially small scale. Since we will also encounter the same difficulty in our nonhyperbolic case, let us explore this idea a little further.

To be more precise, let the coordinates x and y be chosen so that the matrices A and B are in their real Jordan canonical form:

$$A = \text{diag}\,(A_0, A_1) \quad \text{and} \quad B = \text{diag}\,(B_0, B_1)$$

with the property that the real parts for the eigenvalues of the $p \times p$ ($q \times q$, respectively) matrix A_0 (B_0, respectively) are a single number $\lambda_0 < 0$ ($\mu_0 > 0$, respectively), and that those of the $(m-p)\times(m-p)$ ($(n-q)\times(n-q)$, respectively) matrix A_1 (B_1, respectively) are strictly less (greater, respectively) than λ_0 (μ_0, respectively). A_0 (B_0, (respectively) is called the principal stable (unstable, respectively) block and its eigenvalues principal stable (unstable, respectively) eigenvalues. Define

$$A_* = \text{diag}\,(A_0, \lambda_0 I_{(m-p)}) \quad \text{and} \quad B_* = \text{diag}\,(B_0, \mu_0 I_{(n-q)}),$$

where I_i is the $i \times i$ identity matrix. Then, the Sil'nikov solution is said to admit a C^l exponential expansion if it can be expressed as

$$\begin{aligned} x(t) &= e^{A_* t}[\psi(t-\tau, x_0, y_1) + R_1(t, \tau, x_0, y_1)], \\ y(t) &= e^{B_*(t-\tau)}[\varphi(t, x_0, y_1) + R_2(t, \tau, x_0, y_1)] \end{aligned} \tag{1.4a}$$

for all $0 \leqq t \leqq \tau$ and all sufficiently small $|x_0|$ and $|y_1|$ with the properties that the coefficient functions ψ and φ are C^l satisfying

$$\begin{aligned} &\frac{\partial \psi}{\partial x_0}(t-\tau, 0, 0) = \text{diag}\,(I_p, 0), \qquad \frac{\partial \psi}{\partial y_1}(t-\tau, 0, 0) = 0 \quad \text{for all } 0 \leqq t \leqq \tau, \\ &\frac{\partial \varphi}{\partial y_1}(t, 0, 0) = \text{diag}\,(I_q, 0), \qquad \frac{\partial \varphi}{\partial x_0}(t, 0, 0) = 0 \quad \text{for all } t \geqq 0, \end{aligned} \tag{1.4b}$$

and that the remainder terms R_1 and R_2 are also C^l satisfying

$$|D^i R_1(t, \tau, x_0, y_1)| \leqq K e^{-\sigma t}, \qquad |D^i R_2(t, \tau, x_0, y_1)| \leqq K e^{\sigma(t-\tau)}, \tag{1.4c}$$

for all $0 \leqq t \leqq \tau$ and all sufficiently small $|x_0|$ and $|y_1|$, where K and σ are some constants independent of t, τ, x_0, and y_1, and D^i is the ith differentiation operator up to the order $0 \leqq i \leqq l$.

It turns out that a sufficient condition for the exponential expansion requires that the coordinate (x, y) be admissible in the following sense that, besides being of higher order, f has the order $\sum_{k=1}^{p} |x^{(k)}|^2 + \sum_{k=p+1}^{m} |x^{(k)}|$ while g has the order $\sum_{k=1}^{q} |y^{(k)}|^2 + \sum_{k=q+1}^{n} |y^{(k)}|$ as $(x, y) \to (0, 0)$. Note that this necessarily implies that $W^{\text{s}}_{\text{loc}} = \{y = 0\}$ and $W^{\text{u}}_{\text{loc}} = \{x = 0\}$ locally. Fortunately, an admissible coordinate can be obtained by a C^{r-2} change of variables for (1.1), and the exponential expansion is C^{r-4}. For the complete but nontrivial details we refer to Deng (1988a, b, d). A counterexample against the exponential expansion when the coordinate is not admissible is also given in Deng (1988b).

Bearing in mind the questions of what are the admissible variables and what are the corresponding exponential expansions for nonhyperbolic equilibria, let us see what kinds of additional information we can draw from the expansion. First, the local strong unstable manifold $W^{\text{uu}}_{\text{loc}}$ is given by the level set $\varphi(0, 0, y) = 0$ of the expansion coefficient

function. By (1.4b) it can be expressed as the graph of a C^{r-4} function over the last $n-q$ variables $y^{(q+1)}, \cdots, y^{(n)}$. Second, when the system does not have the nonprincipal blocks, the exponential expansion implies that C^1-linearization theorem, constructively (cf. Deng (1988a)). Third, when the principal unstable block is only one-dimensional, we have

$$\lim_{\tau \to +\infty} y^T(0, \tau, x_0, y_1) \cdot \frac{\partial y^T}{\partial \tau}(0, \tau, x_0, y_1)[y^T(0, \tau, x_0, y_1) \cdot y(0, \tau, x_0, y_1)]^{-1} = B_0,$$

which is precisely the theoretical scheme for the convergence of the Feigenbaum number, where a^T means the transpose of a (see, e.g., Collet and Eckmann (1980)). It is my personal belief that this formula also holds true for all finite-dimensional principal blocks with $y(0, \tau, x_0, y_1)$ above being replaced by an $n \times (n-q)$ matrix

$$(y(0, \tau, x_0, y_{1,1}), \cdots, y(0, \tau, x_0, y_{1,n-q})),$$

with the property that the matrix

$$(\varphi(0, x_0, y_{1,1}), \cdots, \varphi(0, x_0, y_{1,n-q}))$$

has the maximal rank q. Last, but not finally, by using the exponential expansion and the Implicit Function Theorem, we can prove the strong λ-lemma, which states that for every point u_0 on the stable manifold there is associated a $(d-n+q)$-dimensional linear space $W(u_0)$, which contains the stable tangent space at u_0 as a subset such that, for every $(n-q)$-dimensional C^{r-7} disc D_0 transverse to this critical affine plane $W(u_0)$, the image D_τ under the time τ map of the flow approaches the strong unstable manifold W^{uu}_{loc} in the C^{r-7}-topology as $\tau \to +\infty$. See Deng (1988d) for a proof. Figure 1.3 illustrates the use of the strong inclination property in classifying some homoclinic

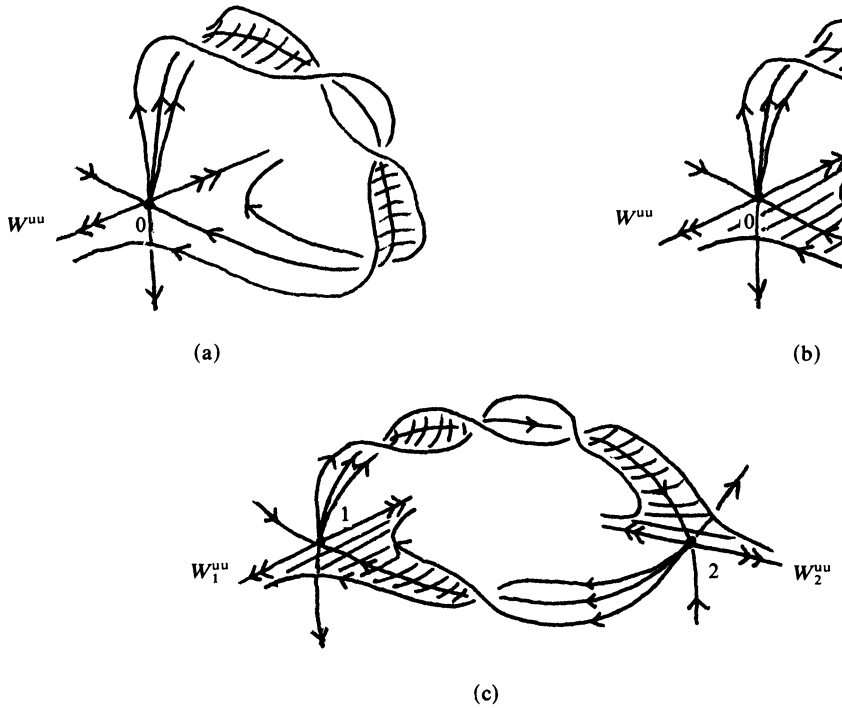


FIG. 1.3. *The phase portaits of some nondegenerate orbits.* (a) *Nontwisted homoclinic orbit.* (b) *Twisted homoclinic orbit.* (c) *Double twisted heteroclinic loop.*

and heteroclinic bifurcations for the flow. What is the strong λ-lemma for nonhyperbolic equilibria, and how can we use it, if at all, to classify homoclinic and heteroclinic orbits? Most important, how can we solve a given homoclinic bifurction problem by combining all these ideas?

We are now in a natural position to outline our paper, giving hints as to the answers. In § 2, specifically in Lemma 2.2, we will use the Center Manifold Theory only to obtain a C^{r-2} admissible coordinate for the system (1.1), with F having additional l eigenvalues of the linearization $DF(0)$ that lie on the imaginary axis of the complex plane. A coordinate $u=(x, y, z)$ is called admissible in this case if in terms of the new variable (1.1) takes the following form:

$$\dot{x}=Ax+f(x, y, z), \quad \dot{y}=By+g(x, y, z), \quad \dot{z}=\theta(z)+h(x, y, z),$$

where f, g, and h are higher-order terms satisfying $f=O(|x|+|y|+|z|)|x|$, $g=O(|x|+|y|+|z|)|y|$, and $h=O(|x||y|)$, and $\dot{z}=\theta(z)$ describes the flow on the local center manifold $\{x=0, y=0\}$ with θ being C^{r-1}. Note that the admissible coordinate directly implies the C^{r-1} "straight" invariant foliations on the center-stable and center-unstable manifolds as $W^{cs}=\bigcup_{|z_0|\ll 1}\{z=z_0, y=0\}$ and $W^{cu}=\bigcup_{|z_0|\ll 1}\{z=z_0, x=0\}$, respectively. In particular, when $z_0=0$, which is W^{u}_{loc}, it is analogous to $W^{uu}_{loc}=\{\varphi(0, 0, y)=0\}$ for the hyperbolic exponential expansion (see Fig. 1.4). The foliations will be very useful in § 5 in establishing the bifurcation equations and the homoclinic and heteroclinic connections between bifurcated equilibria.

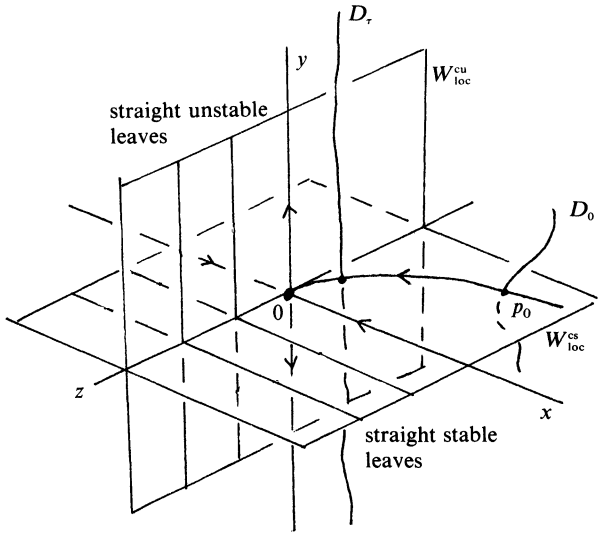


FIG. 1.4. *The straight foliations and the strong λ-lemma.*

In § 3, we will formulate the Sil'nikov solution according to its center flow. Roughly speaking, for every local center flow $z^c(t)$ (i.e., $\dot{z}^c(t)=\theta(z^c(t))$) defined on the positive maximum interval $0\leqq t<\tau^c$ with respect to a fixed small neighborhood of the equilibrium point, there exists a unique local flow $(x, y, z)(t)$ satisfying $x(0)=x_0$, $y(\tau)=y_1$, and $z(0)=z_0=z^c(0)$ (or $z(\tau)=z_1=z^c(\tau)$) for a given triplet (τ, x_0, y_1) with $0\leqq\tau<\tau^c$ and small $|x_0|$ and $|y_1|$. Moreover, this solution can be expanded according to its center flow in the sense that $z(t)=z^c(t)+R(t)$ and the exponential bounds $|D^i x(t)|\leqq K e^{\lambda t}$, $|D^i y(t)|\leqq K e^{\mu(t-\tau)}$, and $|D^i R(t)|\leqq K e^{\lambda t+\mu(t-\tau)}$ are valid for all $0\leqq t\leqq\tau<\tau^c$, and all

sufficiently small $|x_0|$ and $|y_1|$, where $\lambda_0 < \lambda < 0 < \mu < \mu_0$ and K are constants independent of t, τ, τ^c, x_0, y_1, z_0 (or z_1), and the derivatives are taken in t, τ, x_0, y_1, and z_0 (or z_1) up to the orders $i \leqq r-4$. However, the regularity r here must be finite if we want to find those constants. For the precise statement, see Lemma 3.1. The proof is directly based on the uniform contraction mapping principle and has much in common with the existence, uniqueness, and continuous dependence of the Sil'nikov problem (or the initial value problem) for the hyperbolic case. To obtain the exponential bounds, certain weighted Banach spaces are used for functions over $0 \leqq t \leqq \tau$ that are bounded up to some weighted exponential scales—for instance, $e^{-\lambda t}$, $e^{-\mu(t-\tau)}$, and $e^{-\lambda t-\mu(t-\tau)}$ are used for $x(t)$, $y(t)$, and $R(t)$, respectively.

In § 4, we will prove the strong λ-lemma, Lemma 4.1, which is heuristically illustrated in Fig. 1.4. Roughly speaking, it states that if a trajectory on the center-stable manifold approaches the equilibrium point $u = 0$, then for every C^{r-3} n-dimensional disc D_0 that transversely intersects the center-stable manifold through a point of the trajectory, the image D_τ under the time τ mapping of the flow converges to the unstable manifold as $\tau \to +\infty$ in the C^{r-3}-topology. Moreover, the convergence rate is the same as that of the center trajectory, but the tangent space, being normal to the center-stable manifold, "stretches" exponentially rapidly. Note that when the disc D_0 happens to be one of the straight leaves $\{z = z_0, x = 0\}$ on the center unstable manifold, the preceding description makes perfect sense, since in terms of the straight foliation mentioned above, $D_\tau = \{z^c(t, z_0), x = 0\}$ locally.

In § 5, we will first classify nondegenerate homoclinic orbits in general by the strong λ-lemma and just consider three types of nonhyperbolic equilibria in particular: saddle-node, transcritical, and pitchfork. The generic codimension-2 bifurcation unfoldings are obtained through a modified Lyapunov-Schmidt reduction for the Poincaré map in the Sil'nikov variable (see Theorems 5.1–5.3 for the precise statements). One parameter here governs the bifurcations of the equilibria and the other the breaking of the homoclinic orbits. Due to the lack of oscillatory structures for the center flows, all the dynamics considered are nonchaotic. The chaotic bifurcations of a homoclinic orbit to a Hopf equilibrium, or of a transverse homoclinic point to a nonhyperbolic fixed point of a map, are not studied here mainly because many difficulties in analyzing chaos are still under investigation.

As we have seen, the next three sections consist of the foundation of our nonlinear and nonhyperbolic analysis. It allows us to reduce a complex problem simply to an individual case study on the local center manifold. Then the dominant role of the center flow in the bifurcations theory should prevail as usual. Unexpectedly, however, the exponential expansion and the strong λ-lemma for the nonhyperbolic case are much more easily and directly obtained than their hyperbolic counterparts. Of course, to see this we need to compare the proofs with Deng (1988a, d). Also, for the answers that cannot be included in this Introduction, we will refer our readers to Chow, Deng, and Terman (1987), Deng (1988e), and Chow, Deng, and Fiedler (1988) for homoclinic and heteroclinic bifurcation problems with hyperbolic equilibria, which have much in common with the spirit of § 5. For another important topic that is not treated in this paper, we will refer the reader to Schecter (1987) for an example of the saddle-node homoclinic bifurcation in $\mathbb{R}^2$, and to Dangelmayr, Armbruster, and Neveling (1985) and Ju (1988) for an example of the pitchfork homoclinic bifurcation in $\mathbb{R}^2$ as well. The former models the dynamics of the forced Josephson junction, and the latter, the laser with a saturable absorber.

Let us conclude this section with some remarks about our motivations. Luk'yanov (1982) and Schecter (1987) first studied the homoclinic bifurcation with a saddle-node

equilibrium point for planar systems. Chow and Lin (1988) then generalized their results to any finite-dimensional case, using a great variety of techniques, including exponential dichotomy, Melnikov function, smooth foliation, and Sil'nikov's central ideal, called parametrization in their terminology, which also gives rise to the emergence of our method presented here. They found the periodic orbit in a rather geometrical way, not by a Lyapunov-Schmidt reduction technique as we do here. Using their different method, they were the first to realize the necessity of the admissible normal form for the exponential expansion. However, Chow and Lin's technique for the expansion is applicable for only zero-center eigenvalues at the bifurcation point, ruling out the important class of Hopf bifurcation, where the center eigenvalues are nonzero in general. Also, as discussed in Chow, Deng, and Fiedler (1988), to use exponential dichotomy together with the Melnikov function is essentially to ignore the homoclinic doubling bifurcations that are very likely to occur when the center manifold is two-dimensional or the homoclinic orbit is degenerate. Homoclinic doubling bifurcations do occur in some hyperbolic cases (see, e.g., Yanagida (1987) and Chow, Deng, and Fiedler (1988)). Moreover, instead of separated apparatus to the homoclinic and periodic bifurcations, only one bifurcation equation derived from the Lyapunov-Schmidt reduction is needed in our strategy. More important, our main objectives in this paper are to unify the method for homoclinic bifurcation problems regardless of the nature of the equilibria, and to lay the foundation for our future investigations into other, more complicated problems, in particular, chaotical problems.

2. Admissible variables. From now on, we will let $u=(x, y, z)$ with $x\in\mathbb{R}^m$, $y\in\mathbb{R}^n$, $z\in\mathbb{R}^l$, and $m+n+l=d$ such that

$$DF(0)=\operatorname{diag}(A, B, C),$$

where A and B have the same meanings as in § 1 and all the eigenvalues of the $l\times l$ matrix C lie on the imaginary axis of the complex plane. Let W^{cs}, W^{cu}, and W^c denote an $(m+l)$-dimensional center-stable manifold, an $(n+l)$-dimensional center-unstable manifold, and an l-dimensional center manifold, respectively. Then, by the theory of invariant manifolds (see Hirsch, Pugh, and Shub (1977), Vanderbauwhede and van Gils (1987), Wells (1976), and Chow and Lu (1988)), these are C^r manifolds with $r\neq\infty$. Moreover, up to a C^r change of variables, we may assume

$$W^{cs}=\{y=0\},\quad W^{cu}=\{x=0\},\quad W^c=\{x=0, y=0\}$$

locally. Also, when (1.1) is written in terms of such a C^r coordinate, it takes the form

$$\dot{x}=Ax+f(x, y, z),\quad \dot{y}=By+g(x, y, z),\quad \dot{z}=\theta(z)+h(x, y, z) \tag{2.1}$$

with $D\theta(0)=C$ and the nonlinear higher-order terms satisfying

$$\begin{aligned} &f(0, y, z)=0,\quad g(x, 0, z)=0,\quad h(0, 0, z)=0,\\ &Df(0, 0, 0)=0,\quad Dg(0, 0, 0)=0,\quad Dh(0, 0, 0)=0. \end{aligned} \tag{2.1a}$$

Moreover, the functions θ, f, g, and h are C^{r-1}.

DEFINITION 2.1. The coordinate (x, y, z) is admissible if, in addition to condition (2.1a), we have $h(0, y, z)=h(x, 0, z)=0$. A change of variables is admissible if the new variables are admissible.

LEMMA 2.2. *There exists a C^{r-2} admissible change of variables for* (2.1).

Proof. The proof is based on an idea by Ovsyannikov and Sil'nikov (1986) and Deng (1988c), using the Center Manifold Theorem. Let us rewrite (2.1) satisfying

(2.1a) as follows:

$$\begin{aligned} \dot{x} &= Ax + f_1(x, y, z)x, \\ \dot{y} &= By + g_1(x, y, z)y, \\ \dot{z} &= Cz + \bar{\theta}(z) + h_1(x, y, z)x + h_2(x, y, z)y, \end{aligned} \tag{2.2}$$

where $\bar{\theta}(z) = \theta(z) - Cz$ is C^{r-1}, but f_1, g_1, h_1, and h_2 are C^{r-2}. Consider a change of variables

$$x = x, \quad y = y, \quad \zeta = z - p(\zeta, x)x - q(\zeta, y)y$$

with some C^{r-2} functions p and q to be determined satisfying $p(0, 0) = 0$ and $q(0, 0) = 0$. Note that such a change of variables necessarily preserves condition (2.1a). Substituting the new variables ζ into (2.2), we have

$$\begin{aligned} \dot{\zeta} &= \dot{z} - \dot{p}x - p\dot{x} - \dot{q}y - q\dot{y} \\ &= C(\zeta + px + qy) + \bar{\theta}(\zeta + px + qy) + h_1 x + h_2 y - \dot{p}x - p(Ax + f_1 x) - \dot{q}y - q(By + g_1 y), \end{aligned}$$

where h_1, h_2, f_1, and g_1 are understood in the new variables x, y, and ζ. Also, $\dot{p}$ and $\dot{q}$ here are derivatives along the solutions of the new equations. For this reason p and q may also be regarded as variables from $\mathbb{R}^{l\times m}$ and $\mathbb{R}^{l\times n}$, respectively.

Let

$$\bar{\theta}(\zeta + px + qy) - \bar{\theta}(\zeta) = \theta_1(x, y, \zeta, p, q)px + \theta_2(x, y, \zeta, p, q)qy$$

for some C^{r-2} functions θ_1 and θ_2. It is easy to see that

$$\theta_1(0, 0, 0, 0, 0) = \theta_2(0, 0, 0, 0, 0) = 0.$$

Moreover,

$$\theta_1(x, 0, \zeta, p, q) = \theta_1(x, 0, \zeta, p, 0), \qquad \theta_2(0, y, \zeta, p, q) = \theta_2(0, y, \zeta, 0, q). \tag{2.3}$$

Collecting like terms in the equation for $\dot{\zeta}$ above yields

$$\begin{aligned} \dot{\zeta} = C\zeta + \bar{\theta}(\zeta) &+ [Cp + \theta_1 p - \dot{p} - pA - pf_1 + h_1]x \\ &+ [Cq + \theta_2 q - \dot{q} - qB - qg_1 + h_2]y. \end{aligned}$$

Now it is easy to see that, for the new variable to be admissible, it suffices for the first bracket term above to be zero when $y = 0$ and for the second to be zero when $x = 0$. For the first case, this is equivalent to saying that on the center-stable manifold $y = 0$ the following coupled equations must be satisfied:

$$\begin{aligned} \dot{x} &= Ax + f(x, 0, \zeta + px), \\ \dot{\zeta} &= C\zeta + \bar{\theta}(\zeta), \\ \dot{p} &= Cp - pA + \theta_1(x, 0, \zeta, p, q)p - pf_1(x, 0, \zeta + px) + h_1(x, 0, \zeta + px). \end{aligned} \tag{2.4}$$

Note that these equations do not actually depend on the q variable since $\theta_1(x, 0, \zeta, p, q) = \theta_1(x, 0, \zeta, p, 0)$ according to (2.3). The linearization of this vector field of $(m + l + ml)$ equations at the trivial equilibrium point of the origin has a lower triangular form whose diagonal blocks consist of the stable matrix A, the center matrix C, and the matrix for the linear operator $Lp = Cp - pA$ for all $l \times m$ matrices p. Thus the set of eigenvalues consists of $\Sigma(A)$, $\Sigma(C)$, and $\Sigma(L)$, where $\Sigma(A)$ is the set of eigenvalues of a given linear operator A. Let us determine $\Sigma(L)$. It is easy to check directly that if λ is an eigenvalue for the transpose matrix A^* and v is a corresponding eigenvector, and likewise, if $\mu \in \Sigma(C)$ with a corresponding eigenvector w, then wv^* is an eigenvector of L for the eigenvalue $\mu - \lambda$, whose real parts are positive for all $\lambda \in \Sigma(A)$ and $\mu \in \Sigma(C)$. Moreover, $\mu - \lambda$ are the only eigenvalues, since the dimension of the generalized eigenvector space corresponding to $\mu - \lambda$ is the product of those of λ and μ (see, e.g., Lancaster (1969)).

Because of such a separation of the eigenvalues, the theory of invariant manifolds (see the same references above) applies. Thus, there exists a C^{r-2} function $p = p(\zeta, x)$ whose graph gives rise to the center-stable manifold of this (x, ζ, p) system of (2.4). The same argument yields the function q. □

Remark 2.3. (a) If (1.1) is C^r differentiably depending on a parameter, then the admissible change of variables also smoothly varies with the parameter. This can be directly achieved by the lemma, treating the parameter as an additional center flow.

(b) Under the admissible variables, the function θ in (2.1) remains unchanged and thus is C^{r-1}, but f and g are reduced to C^{r-2} and h is C^{r-3}.

(c) As mentioned earlier in the Introduction, we obtain the "straight" invariant foliations $W^{cs} = \bigcup_{|z_0| \ll 1} \{z = z_0, y = 0\}$ and $W^{cu} = \bigcup_{|z_0| \ll 1} \{z = z_0, x = 0\}$ as a corollary to the admissible change of variables. In particular, the local stable and local unstable manifolds are, respectively, the x-axis ($y = 0$, $z = 0$) and the y-axis ($x = 0$, $z = 0$) locally. For different approaches to achieve the same foliation there exists a geometric proof based on the graph transformation method by Hirsch, Pugh, and Shub (1977) extensively for diffeomorphisms and Fenichel (1979) for flows, and an analytic proof based on the variation of constants formula by Henry (1981) and later by Chow, Lin, and Lu (1988). In contrast to our approach, the admissible variables can also be obtained through their invariant foliations.

3. Exponential expansion with center flows. Let $z^c(t)$ with $z^c(0) = z_0$ (or $z^c(\tau) = z_1$) be any solution on the center manifold defined for $0 \leqq t < \tau^c$ with respect to a certain neighborhood of the origin, where τ^c could be infinity. Given such a center solution and a triplet (τ, x_0, y_1) with $0 \leqq \tau < \tau^c$, a solution $(x, y, z)(t)$ of (2.1) is called a Sil'nikov solution if the Sil'nikov conditions $x(0) = x_0$, $y(\tau) = y_1$, and $z(0) = z^c(0)$ are satisfied. This is sometimes referred to as the first type of Sil'nikov problem. The second type of Sil'nikov problem is, of course, the same as the first one except that the last condition $z(0) = z^c(0)$ is replaced by $z(\tau) = z^c(\tau)$. Indeed, they are identical up to the time reversal $(t \to -t)$. Suppose that the Sil'nikov solution exists and is unique with respect to the Sil'nikov conditions for all $0 \leqq \tau < \tau^c$ and sufficiently small $|x_0|$, $|y_1|$, and $|z_0|$ (or $|z_1|$) and that the function $(x, y, z)(t) \overset{\text{def}}{=} (x, y, z)(t, \tau, x_0, y_1, z_0)$ (or $(x, y, z)(t, \tau, x_0, y_1, z_1)$) is C^k for all the arguments. Then the solution is said to admit an exponential expansion of regularity k if there exists a C^k function R of (t, τ, x_0, y_1, z_0) (or (t, τ, x_0, y_1, z_1)) such that the following is satisfied:

$$z(t) = z^c(t) + R(t, \tau, x_0, y_1, z_0) \quad (\text{or } R(t, \tau, x_0, y_1, z_1)) \tag{3.1a}$$

with

$$R(0, \tau, x_0, y_1, z_0) = 0 \quad (\text{or } R(\tau, \tau, x_0, y_1, z_1) = 0), \tag{3.1b}$$

and there exist constants $\lambda_0 < \lambda < 0 < \mu < \mu_0$ and K independent of t, τ, x_0, y_1 and z_0 (or z_1) such that

$$|D^i x(t)| \leqq K e^{\lambda t}, \quad |D^i y(t)| \leqq K e^{\mu(t-\tau)}, \quad |D^i R(t)| \leqq K e^{\lambda t + \mu(t-\tau)} \quad \text{for } 0 \leqq t \leqq \tau, \tag{3.1c}$$

where D^i denotes the ith derivative in all the arguments up to the order $0 \leqq i \leqq k$.

For definiteness, let us consider the first type of Sil'nikov problem in the following lemma. Necessary modifications for the second type are given in the remarks after the proof.

LEMMA 3.1. *Let the variable (x, y, z) for (2.1) be C^{r-2} admissible as in Lemma 2.2. Let $\beta > 0$, $\lambda < 0 < \mu$ be arbitrary but fixed constants satisfying $\lambda_0 + \beta(r-2) < \lambda < 0 < \mu < \mu_0 - \beta(r-2)$ and $\lambda + \mu - \beta(r-2) > 0$. Then there exist positive constants M, K, and small δ_0 depending on the choices of β, λ, and μ only such that as long as $|z^c(t)| \leqq \delta_0$*

for $0 \leqq t < \tau^c$, *there exists a unique Sil'nikov solution for all* $0 \leqq \tau < \tau^c$, $|x_0|$, $|y_1|$, *and* $|z_0| \leqq \delta_0$, *which admits a* C^{r-3} *exponential expansion. In particular, for the solution itself the constant* K *in* (3.1c) *can be replaced by* $2M\delta_0$.

Proof. The proof is based on the uniform contraction mapping principle. Let

$$\sigma = \min\{\lambda - \lambda_0 - \beta(r-2), \mu_0 - \mu - \beta(r-2), \lambda + \mu - \beta(r-2)\}$$

and M be large enough so that

$$|e^{At}| \leqq M e^{(\lambda_0+\beta)t} \quad \text{for } t \geqq 0, \qquad |e^{Bt}| \leqq M e^{(\mu_0-\beta)t} \quad \text{for } t \leqq 0,$$
$$|e^{Ct}| \leqq M e^{\beta|t|} \quad \text{for all } t.$$

Let $\|f\|$ be the C^{r-2} norm of a given function f in the neighborhood of the origin for which the admissible form (2.1) is valid. Let

$$\delta_0 = \left[\frac{4M^2}{\sigma}(\|f\| + \|g\| + \|\theta\| + \|h\|)\right]^{-1}. \tag{3.2}$$

Let $R(t) = z(t) - z^c(t)$. Let us consider the equations for x, y, and R. We have

$$\dot{R}(t) = \dot{z}(t) - \dot{z}^c(t) \stackrel{\text{def}}{=} CR(t) + L(z^c(t), R(t))R(t) + h(x(t), y(t), R(t) + z^c(t)),$$

where

$$L(z^c, R) = \int_0^1 \frac{d\bar{\theta}}{dz}(sR + z^c)\, ds$$

with $\bar{\theta}(z) = \theta(z) - Cz$ as in (2.1). Now, it is easy to check that the existence of the Sil'nikov solution is equivalent to the existence of the solutions to the following integral equations:

$$\begin{aligned} x(t) &= e^{At}x_0 + \int_0^t e^{A(t-s)} f(x, y, R + z^c)\, ds, \\ y(t) &= e^{B(t-\tau)}y_1 + \int_\tau^t e^{B(t-s)} g(x, y, R + z^c)\, ds, \\ R(t) &= \int_0^t e^{C(t-s)}[L(z^c, R)R + h(x, y, R + z^c)]\, ds, \end{aligned} \tag{3.3}$$

where $x = x(s)$, $y = y(s), \cdots$, etc., in the integrals. Let Σ be the set of continuous functions $(x, y, R)(t)$ defined on $0 \leqq t \leqq \tau < \tau^c$ satisfying $|x(t)| \leqq 2\delta_0 M e^{\lambda t}$, $|y(t)| \leqq 2\delta_0 M e^{\mu(t-\tau)}$, and $|R(t)| \leqq 2\delta_0 M e^{\lambda t + \mu(t-\tau)}$. Equip Σ with a weight norm

$$\|(x, y, R)\|_\Sigma = \sup_{0 \leqq t \leqq \tau} (|x(t)\, e^{-\lambda t}| + |y(t)\, e^{-\mu(t-\tau)}| + |R(t)\, e^{-\lambda t - \mu(t-\tau)}|).$$

Let T denote the operator defined by the right-hand side of the integral equation (3.3). Then T is a contraction mapping on Σ with the contraction constant

$$\rho = \frac{2M^2}{\sigma}(\|f\| + \|g\| + \|\theta\| + \|h\|)\delta_0 = \frac{1}{2}.$$

Indeed, for $(x, y, R) \in \Sigma$ and $(\bar{x}, \bar{y}, \bar{R}) = T(x, y, R)$, we have

$$\begin{aligned} |\bar{x}(t)| &\leqq \delta_0 M e^{\lambda t} + \int_0^t M e^{(\lambda-\sigma)(t-s)} 4M^2\delta_0^2 \|f\|\, e^{\lambda s}\, ds \\ &\leqq \delta_0 M e^{\lambda t} + \frac{4M^3\|f\|}{\sigma}\delta_0^2 e^{\lambda t} \leqq 2\delta_0 M e^{\lambda t}, \end{aligned}$$

because of (3.2) and $|f| = O(|x|+|y|+|z|)|x|$ and $\lambda_0+\beta \leqq \lambda-\sigma$. Similarly, we have $|\bar{y}| \leqq 2\delta_0 M e^{\mu(t-\tau)}$. Moreover,

$$|\bar{R}(t)| \leqq \int_0^t M e^{\beta(t-s)}[4M^2\|\theta\|\delta_0^2 e^{\lambda s+\mu(s-\tau)}]\,ds,$$

since $|L| = O(|R|+|z^c|)$, $|h| = O(|x\|y|)$. Thus,

$$|\bar{R}(t)| \leqq \frac{4M^3(\|\theta+\|h\|)\delta_0^2}{\lambda+\mu-\beta} e^{\lambda t+\mu(t-\tau)}$$
$$\leqq 2\delta_0 M e^{\lambda t+\mu(t-\tau)}$$

because of (3.2) and $\lambda+\mu-\beta \geqq \sigma$. Hence, $T(\Sigma) \subset \Sigma$. To show T is contractive, observe the following trivial estimates:

$$|f(x, y, R+z^c)-f(\tilde{x}, \tilde{y}, \tilde{R}+z^c)| \leqq 2M\delta_0\|f\| e^{\lambda t}\|(x, y, R)-(\tilde{x}, \tilde{y}, \tilde{R})\|_\Sigma,$$
$$|g(x, y, R+z^c)-g(\tilde{x}, \tilde{y}, \tilde{R}+z^c)| \leqq 2M\delta_0\|g\| e^{\mu(t-\tau)}\|(x, y, R)-(\tilde{x}, \tilde{y}, \tilde{R})\|_\Sigma$$

for (x, y, R) and $(\tilde{x}, \tilde{y}, \tilde{R}) \in \Sigma$, since $|f| = O(|x|+|y|+|z|)|x|$ and $|g| = O(|x|+|y|+|z|)|y|$. Also

$$|L(z^c, R)R+h(x, y, R+z^c)-(L(z^c, \tilde{R})\tilde{R}+h(\tilde{x}, \tilde{y}, \tilde{R}+z^c))|$$
$$\leqq 2M\delta_0(\|\theta\|+\|h\|) e^{\lambda t+\mu(t-\tau)}\|(x, y, k)-(\tilde{x}, \tilde{y}, \tilde{R})\|_\Sigma$$

since $h(x, y, z) = O(|x||y|)$. Now, similarly to the estimates for $\bar{x}$, $\bar{y}$, and $\bar{R}$ above, it is easy to check that

$$\|(\bar{x}, \bar{y}, \bar{R})-(\bar{\tilde{x}}, \bar{\tilde{y}}, \bar{\tilde{R}})\|_\Sigma \leqq \rho\|(x, y, R)-(\tilde{x}, \tilde{y}, \tilde{R})\|_\Sigma$$

with $\rho = \frac{1}{2}$ as above. Thus, by the Uniform Contraction Mapping Theorem, there exists a unique function ζ^* in Σ such that $\zeta^* = T(\zeta^*)$. Moreover, $\zeta^*(t) \overset{\text{def}}{=} \zeta^*(t, \tau, x_0, y_1, z_0)$ is C^{r-3} in the parameters x_0, y_1, and z_0 since the admissible change of variables is C^{r-2}. To show it is also C^{r-3} at $\tau = \tau_0 < \tau^c$, we simply replace the interval $[0, \tau_0]$ on which all the functions of the space Σ are defined by a larger one $[0, \tau_0+\varepsilon]$ and then show that the same operator T has a unique fixed point $\zeta^*(t, \tau, x_0, y_1, z_0)$ in the new function space. Thus $\zeta^*(t, \tau, x_0, y_1, z_0)$ is C^{r-3} in τ_0 as well. The differentiability in time t simply follows the standard argument for the smoothness of solutions to the initial value problem found in textbooks, for instance, Hale (1978) and Irwin (1980). Since all the partial derivatives are continuous, ζ^* is C^{r-3} differentiable by a standard fact from calculus.

The estimates for the derivatives follow the same technique as for the exponential bounds above. To begin with, we observe first that the growth rate for all the variational flow $D^j z^c(t, z_0)$ on the center manifold cannot be greater than $O(e^{j\beta|t|})$ for all $0 \leqq j \leqq r-3$ and all t as defined. Indeed, this can be proved directly by using the same arguments as above, using an appropriate weight norm for all the functions $z(t)$, i.e., the maximum of the exponentially scaled function $|e^{-j\beta t}z(t)|$ over $0 \leqq t < \tau^c$ (see also, e.g., Vanderbauwhede and van Gils (1987)). When these estimates are used for the variational integral equations for the mixed system obtained by differentiating (3.3), we clearly see that the corresponding derivatives on x, y, and R will not exceed the orders $e^{(\lambda_0+(j+1)\beta)t}$, $e^{(\mu_0-(j+1)\beta)(t-\tau)}$, and $e^{(\lambda_0+(j+1)\beta)t+(\mu_0-(j+1)\beta)(t-\tau)}$, respectively. Again, the desired estimates are obtained by the uniform contraction mapping principle together with some appropriate weighted Banach spaces of functions. We note that all the contraction constants are the same number $\rho = \frac{1}{2}$ but the constant $2M\delta_0$ may vary for each $0 \leqq j \leqq r-3$. By the choices of λ, μ, β and of a sufficiently large K (since r is finite), the proof is completed. □

Remark 3.2. (a) For the second type of Sil'kinov problem with $z(\tau) = z^c(\tau) = z_1$, Lemma 3.1 is still valid by changing the inequality $\lambda + \mu - \beta(r-2) > 0$ into $\lambda + \mu + \beta(r-2) < 0$. This is obtained by directly applying Lemma 3.1 to the time-reversed system. We will actually use this second type of lemma in § 5.

(b) Later we will also use the proven fact that the variational flow $D^j z^c(t, z_0)$ cannot grow faster than $O(e^{j\beta t})$ for all $0 \leqq t < \tau^c$ and $0 \leqq j \leqq r-3$. More precisely, $|D^j z^c(t, z_0)| \leqq K e^{j\beta t}$ for $0 \leqq t < \tau^c$, where the constant K may be chosen as the same one in the lemma.

(c) All the results above can also be easily extended to systems depending on parameters by the same modification as in § 2 (see Remark 2.3(a))—that is, by considering the parameters as additional center directions.

4. Strong λ-lemma. We will continue to use the same notation and results from the two previous sections. In this section we consider the inclination behaviors of subsets when carried by the local flow forward in time. To be precise, let D^n be an n-dimensional C^{r-3} manifold intersecting the center-stable manifold W^{cs} transversely at a point $p_0 = (x_0, 0, z_0)$. Thus D^n can be written as the graph of a C^{r-3} function $(x, z) = (p, q)(y)$ over a small δ-box $B^n(\delta) = \{y \,|\, |y| \leqq \delta\}$ on the y-axis. Let D^n_τ denote the connected image of D^n in the δ_0-box $B^d(\delta_0)$ of the origin under the time τ mapping of the flow. We are interested in the asymptotic inclination behavior of D^n_τ as $\tau \to +\infty$ under the assumption that the center-stable trajectory through the point of intersection p_0 converges to the origin as $\tau \to +\infty$. As observed in the introduction, when D^n happens to be one of the straight leaves $\{z = z_0, x = 0\}$ on the center-unstable manifold, the asymptotic inclination behavior is self-evident: D^n converges to the unstable manifold, with the tangent spaces identically equal to each other, so long as the center trajectory goes to the origin (cf. Fig. 1.4). But, in general, D^n_τ is said to be C^{r-3} ε-close to the unstable manifold W^u_{loc} in $B^d(\delta_0)$ by an arbitrarily small number $\varepsilon > 0$ if there exists a time $\tau_*(\varepsilon)$ such that for every $\tau \geqq \tau_*(\varepsilon)$, D^n_τ is the graph of a C^{r-3} function $(x, z) = (p_\tau, q_\tau)(y)$ over $|y| \leqq \delta_0$ satisfying $\|(p_\tau, q_\tau)\| < \varepsilon$, where $\|\cdot\|$ denotes the usual C^{r-3} norm for all C^{r-3} functions in $B^n(\delta_0)$. Note that δ_0 is fixed but $0 < \delta \leqq \delta_0$ is not, in general. Now we have Lemma 4.1.

LEMMA 4.1 (strong λ-lemma). *Given an n-dimensional C^{r-3} disc D^n transversely intersecting the center-stable manifold at a point p_0, if the solution through p_0 converges to the origin as $\tau \to +\infty$, then the image D^n_τ in $B^d(\delta_0)$ under the time τ-mapping of the flow is C^{r-3} ε-close to the unstable manifold. Moreover, the tangent space of D^n_τ at p_τ is exponentially close to the tangent unstable manifold at the origin.*

Proof. The idea of the proof is to use the Sil'nikov solution to express D^n_τ, then the Implicit Function Theorem together with the exponential expansion to obtain the graph representation of D^n_τ, and last, the expansion to estimate the rate of convergence.

Without loss of generality, we may assume that the center trajectory through the projection point $(0, 0, z_0) \in W^c_{loc}$ of the base point $p_0 = (x_0, 0, z_0)$ converges to the origin forward in time. Indeed, this fact simply follows the straight invariant foliation on the center-stable manifold due to the admissible variable and the assumption. Thus, let us assume $|z^c(t, z_0)| \leqq \delta_0/4$ for all $t \geqq 0$.

By definition,

$$D^n_\tau = \{(x_1, y_1, z_1) \,|\, (x_1, y_1, z_1) = (x, y, z)(\tau, 0, p(y_0), y_0, q(y_0)) \text{ for those } |y_0| \leqq \delta \text{ so that } |x_1|, |y_1| \text{ and } |z_1| \leqq \delta_0\}.$$

To use the Sil'nikov solution for the desired alternative representation of D^n_τ, we need to estimate first the definition time for the center flow. Since $|z^c(t, q(0))| \leqq \delta_0/4$ with

$q(0)=z_0$ for all $t\geqq 0$, by the continuous dependence on the initial data we have that for every $\tau>0$ there exists a small number $\gamma(\tau)$ such that $|z^c(t, q(y_0))|\leqq \delta_0/2$ for all $|y_0|\leqq\gamma(\tau)$ and $0\leqq t\leqq\tau$. In fact, we can obtain a better approximation $\gamma(\tau)=(\delta_0/4K\|q\|)\,e^{-\beta\tau}$ by the following estimate:

$$\begin{aligned}|z^c(t, q(y_0))|&\leqq|z^c(t, q(y_0))-z^c(t, z_0)|+|z^c(t, z_0)|\\&\leqq K e^{\beta t}\|q\|\gamma(\tau)+\frac{\delta_0}{4}\leqq\frac{\delta_0}{2}\end{aligned}$$

provided $t\leqq\tau$. Thus, the exponential expansion implies that for $|y_0|\leqq\gamma(\tau)$

$$y_0=y(0, \tau, p(y_0), y_1, q(y_0))$$

holds true provided $|y_1|\leqq\delta_0$. By comparing the two sides of this formula we can easily show by the exponential bounds (3.1c) with $K=2M\delta_0$ when $i=0$ that for $\tau\geqq\tau_1\overset{\text{def}}{=}1/(\mu-\beta)\ln(8MK\|q\|)$, the relation $|y_1|\leqq\delta_0$ must imply $|y_0|<\gamma(\tau)$. Thus, in terms of the Sil'nikov solution, D_τ^n can be written as

$$\begin{aligned}D_\tau^n=\{(x_1, y_1, z_1)\,|\,&x_1=x(\tau, \tau, p(y_0), y_1, q(y_0)), y_0=y(0, \tau, p(y_0), y_1, q(y_0)),\\&z_1=z(\tau, 0, p(y_0), y_0, q(y_0)), \text{for those } |y_0|\leqq\gamma(\tau) \text{ such that } |x_1|, |y_1|, |z_1|\leqq\delta_0\}.\end{aligned}$$

To express D_τ^n as the graph of a C^{r-3} function over $|y_1|\leqq\delta_0$, we use the Implicit Function Theorem to solve the equation

$$\Psi_\tau(y_0, y_1)\overset{\text{def}}{=}y_0-y(0, \tau, p(y_0), y_1, q(y_0))=0$$

for y_0 in terms of y_1. Since $\Psi_\tau(0, 0)=0$ and the Jacobian

$$\left|\frac{\partial\Psi_\tau}{\partial y_0}\right|\geqq 1-\left|\frac{\partial}{\partial y_0}y(0, \tau, p(y_0), y_1, q(y_0))\right|\geqq 1-K(\|p\|+\|q\|)\,e^{-\mu\tau}>\frac{1}{2}$$

for $\tau\geqq\tau_2\overset{\text{def}}{=}(1/\mu)\ln(2K(\|p\|+\|q\|))$ and all $|y_1|\leqq\delta_0$ and $|y_0|\leqq\delta$, we can solve $y_0=\psi_\tau(y_1)$ from the equation for sufficiently small $|y_1|$. Moreover, $|\psi_\tau(y_1)|\leqq 2M\delta_0\,e^{-\mu\tau}$. Note that the last inequality actually implies that the domain of the solution ψ_τ can be extended to the entire δ_0-box, while still maintaining the constraint $|y_0|\leqq\gamma(\tau)$ for all $\tau\geqq\tau_1+\tau_2$. Furthermore, $\|\psi_\tau\|\leqq 2K\,e^{-\mu\tau}$. Let $p_\tau(y_1)=x(\tau, \tau, p(\psi_\tau(y_1)), y_1, q(\psi_\tau(y_1)))$ and $q_\tau(y_1)=z(\tau, 0, p(\psi_\tau(y_1)), \psi_\tau(y_1), q(\psi_\tau(y_1)))$ over $|y_1|\leqq\delta_0$. This completes the second part of the proof.

Last, let us estimate the rate of convergence. It is obvious that we have $\|p_\tau\|=O(e^{\lambda\tau})<\varepsilon$ for large τ. Moreover, by the expansion and Remark 3.2(b) on the growth rate of the center flows,

$$\begin{aligned}|q_\tau(y_1)|&\leqq|z^c(\tau, q(\psi_\tau(y_1)))|+|R(\tau, \tau, p(\psi_\tau(y_1)), y_1, q(\psi_\tau(y_1)))|\\&\leqq|z^c(\tau, q(0))|+\left\|\frac{\partial z^c}{\partial z_0}(\tau, \cdot)\right\|\|q\||\psi_\tau(y_1)|+2M\delta_0\,e^{\lambda\tau}\\&\leqq\frac{\varepsilon}{2}+K\,e^{(r-3)\beta\tau}\|q\|\,e^{-\mu\tau}+2M\delta_0\,e^{\lambda\tau}\\&\leqq\frac{\varepsilon}{2}+\frac{\varepsilon}{4}+\frac{\varepsilon}{4}<\varepsilon\end{aligned}$$

for sufficiently large τ, provided $|z^c(\tau, q(0))|<\varepsilon/2$. The last inequality is true since $z^c(t, q(0))\to 0$ as $t\to+\infty$. Finally, since $z^c(\tau, q(0))$ does not depend on y_1, all the derivatives for p_τ and q_τ in y_1 up to the order $r-3$ are exponentially small. □

As we know, center-stable and center-unstable manifolds are not necessarily unique. However, we have the following corollary that will also be used in § 5.

COROLLARY 4.2. *Two given center-stable manifolds have the same tangent space at any common point whose trajectory converges to the equilibrium.*

Proof. Let W_1^{cs} and W_2^{cs} be two center-stable manifolds intersecting at a point p. If they do not have the same tangent space at p, then there must be at least one tangent direction, say v, of W_1^{cs} normal to W_2^{cs}. Using W_2^{cs} as the center-stable manifold in Lemma 4.1, the limiting direction of v at the origin must be contained in the tangent unstable manifold at the origin that is normal to the center eigenvector space. This is a contradiction since both W_1^{cs} and W_2^{cs} have the same tangent space—the center eigenvector space—at the origin. □

5. Homoclinic bifurcations with nonhyperbolic equilibria. In this section we will classify homoclinic orbits with nonhyperbolic equilibria according to the strong inclination property from the previous section, and consider specifically three basic types of codimension-1 nonhyperbolic equilibria that undergo the saddle-node, transcritical, and pitchfork bifurcations, respectively. We will state and prove the corresponding theorems for the generic two-parameter unfoldings.

For definiteness, from now on we will explicitly assume that the vector field $F = F(u, \alpha)$ of (1.1) depends on two parameters $\alpha = (\alpha_1, \alpha_2)$ in the C^r fashion. Also, for directness we will assume the parameter is generic in the sense that α_1 governs the bifurcations of the equilibria while α_2 governs the transverse crossing of the center-unstable manifold and the stable manifold. This will be made precise as we proceed.

In this paper, we consider only the bifurcation of those homoclinic orbits at the bifurcation point $\alpha = 0$, $\Gamma = \Gamma(t)$ that are contained in either $W^{cu} \cap W^s$ or $W^{cs} \cap W^u$. Up to the time reversal ($t \to -t$) we will always assume the first case. We will also assume that the homoclinic orbit is asymptotically tangent to the center eigenvector space of the linearization $DF(0, 0)$ as $t \to -\infty$. A homoclinic orbit Γ satisfying these two conditions is called nondegenerate if, in addition, there exists an n-dimensional C^1 disc D^n on the center-unstable manifold such that as a submanifold of W^{cu} it transversely intersects the center manifold, while as a submanifold of $\mathbb{R}^d$ it transversely intersects the center-stable manifold W^{cs} at a point from Γ. Observe that when the center dimension l is 1, the nondegeneracy of a Γ is equivalent to the transverse intersecting of the center-unstable and center-stable manifolds along Γ, i.e.,

$$\text{span}\{T_p W^{cs}, T_p W^{cu}\} = \mathbb{R}^d \quad \text{for all } p \in \Gamma, \tag{5.1a}$$

where $T_p W$ means the tangent space of a smooth manifold W at a base point p (see Fig. 5.1). Also note that, since the tangent spaces of center-stable and center-unstable manifolds are uniquely defined along a homoclinic orbit Γ by Corollary 4.2, the definition of nondegeneracy is independent of the choices of these manifolds (for a different justification, see Chow and Lin (1988)).

Let Σ be any $(d-1)$-dimensional small and closed Poincaré cross section transverse to the homoclinic orbit Γ. Let $W^{cu}(\alpha)$ and $W^s(\alpha)$ denote the parametrically dependent center-unstable and stable manifolds that also vary with the parameter α in the C^r fashion. Let $d(\alpha_1, \alpha_2)$ be the distance between $\Sigma \cap W^{cu}(\alpha)$ and $\Sigma \cap W^s(\alpha)$, where d is continuous and $d(0, 0) = 0$, which represents the existence of the original homoclinic orbit Γ. The crossing of the center-unstable and stable manifolds is said to be transverse along the α_2-direction if the following condition is satisfied:

$$\lim_{\alpha_2 \to 0} \frac{d(0, \alpha_2)}{|\alpha_2|} \neq 0. \tag{5.1b}$$

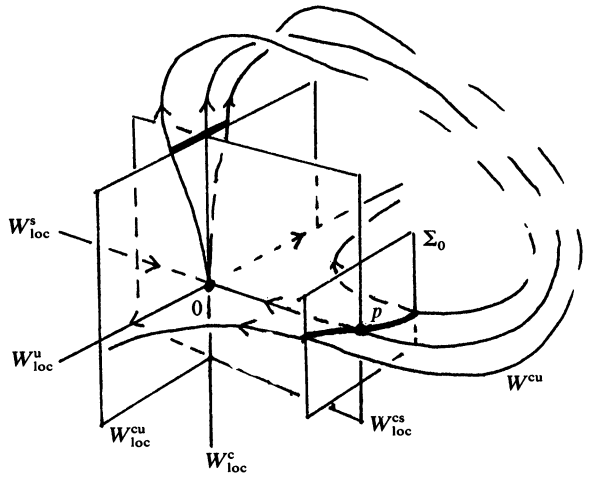


FIG. 5.1. *The phase portrait of a nondegenerate homoclinic orbit.*

Since the flow from one Poincaré cross section to another gives rise to a diffeomorphism, this property of nonzero limit, and consequently the definition of transverse crossing above, are independent of the choices of cross sections.

Next, let us introduce the types of codimension-1 bifurcations of the equilibria to be considered. Let us assume that the linearization $D_uF(0, 0)$ has only one eigenvalue with zero real part, and that the equilibrium point $u = 0$ at $\alpha = 0$ is any of the following:

(5.2a) Saddle-node: $e_l D_u^2 F(0,0)(e_r, e_r) > 0$, $e_l D_{\alpha_1} F(0,0) > 0$, $e_l D_{\alpha_2} F(0,0) = 0$,

(5.2b) Transcritical: $F(0, \alpha) = 0$ for all α, $e_l D_u^2 F(0,0)(e_r, e_r) > 0$, $e_l D^2_{u\alpha_1} F(0,0) e_r > 0$, $e_l D^2_{u\alpha_2} F(0,0) e_r = 0$,

(5.2c) Pitchfork: $F(0, \alpha) = 0$, $e_l D_u^2 F(0, \alpha)(e_r, e_r) = 0$ for all α, $e_l D_u^3 F(0,0)(e_r, e_r, e_r) > 0$, $e_l D^2_{u\alpha_1} F(0,0) e_r > 0$, $e_l D^2_{u\alpha_2} F(0,0) e_r = 0$,

where e_r and e_l are a right and a left eigenvector of the zero eigenvalue, respectively, with e_l chosen so that $e_l e_r > 0$ (see Sotomayor (1974), Guckenheimer and Holmes (1983)). In terms of the center manifold, these intimidating and technical conditions can always be reinterpreted, respectively, in the following relatively explicit ways. Indeed, if we let $\theta = \theta(z, \alpha)$ be the vector field on the center manifold as in (2.1) of § 2, then we have, equivalent to (5.2a-c), respectively,

(5.3a) $$\theta(0,0) = \frac{\partial \theta}{\partial z}(0,0) = 0, \quad \frac{\partial^2 \theta}{\partial z^2}(0,0) > 0, \quad \frac{\partial \theta}{\partial \alpha_1}(0,0) > 0, \quad \frac{\partial \theta}{\partial \alpha_2}(0,0) = 0,$$

(5.3b) $$\theta(0, \alpha) = 0 \text{ for all } \alpha, \quad \frac{\partial \theta}{\partial z}(0,0) = 0, \quad \frac{\partial^2 \theta}{\partial z^2}(0,0) > 0,$$
$$\frac{\partial^2 \theta}{\partial z\, \partial \alpha_1}(0,0) > 0, \quad \frac{\partial^2 \theta}{\partial z\, \partial \alpha_2}(0,0) = 0,$$

(5.3c) $$\theta(0, \alpha) = \frac{\partial^2 \theta}{\partial z^2}(0, \alpha) = 0 \text{ for all } \alpha, \quad \frac{\partial \theta}{\partial z}(0,0) = 0,$$
$$\frac{\partial^3 \theta}{\partial z^3}(0,0) > 0, \quad \frac{\partial^2 \theta}{\partial z\, \partial \alpha_1}(0,0) > 0, \quad \frac{\partial^2 \theta}{\partial z\, \partial \alpha_2}(0,0) = 0.$$

We emphasize once again the explicit roles forced on the parameters as in (5.1b) and (5.2a–c) are simply for definiteness and they can be achieved by following the procedure below. Take the first case (5.2a) as an example. $e_1 D_{\alpha_1} F(0,0) > 0$ and $e_1 D_{\alpha_2} F(0,0) = 0$ in (5.2a) can always be obtained by choosing α_1 as the gradient direction of the scalar function $e_1 F(0, \cdot)$ and α_2 the normal direction to the gradient vector at $\alpha = 0$. Once this is done, it only remains to check condition (5.1b) if an application problem ever arises. As another remark, let us point out that the last two bifurcations of steady states are not generically of codimension 1. They can always be perturbed into a saddle-node equilibrium point by making $\partial\theta/\partial\alpha_1(0,0) \neq 0$. But they do appear in many applications due to other mechanisms, e.g., certain types of symmetries adhering to the physical models considered will force the persistence of the transcritical or pitchfork steady states. See Guckenheimer and Holmes (1983), and in particular, Dangelmayr, Armbruster, and Neveling (1985) and Ju (1988) for specific examples. Nevertheless a two-parameter family of vector fields having a nondegenerate homoclinic orbit to a nonhyperbolic equilibrium point (of the preceding three types) is said to be generic (in our restrictive sense above and in this paper only) if up to a C^r change of parameters the transverse crossing condition (5.1b) and one of the three nonhyperbolic conditions (5.2a–c) are satisfied.

We are now in a position to state our main theorems. Before doing so, we discuss preliminary results on the local bifurcations of steady states as preparation. For the original account of those results, see Sotomayor (1974). We should be aware that all the discussions are valid only in an implicitly small but fixed neighborhood of the origin $(u, \alpha) = (0,0)$ in $\mathbb{R}^{d+2}$. Let us begin with the saddle-node case.

Solving the equation $\theta(z, \alpha) = 0$ and $\partial\theta/\partial z(z, \alpha) = 0$ simultaneously for z and α_1 by the Implicit Function Theorem (IFT), we obtain the continuation of the saddle-node equilibrium points $z = E_0(\alpha_2)$ along a curve $\alpha_1 = c_0(\alpha_2)$. Both functions are C^{r-1} and satisfy

$$E_0(0) = c_0(0) = E_0'(0) = c_0'(0) = 0, \tag{5.4a}$$

because of $\partial\theta/\partial\alpha_2(0,0) = 0$. To find hyperbolic equilibria near $u = 0$ we solve the equation $\theta(z, \alpha) = 0$ alone this time for α_1 by the IFT and obtain a C^{r-1} function $\alpha_1 = \gamma(z, \alpha_2)$ satisfying

$$\begin{aligned} &\gamma(E_0(\alpha_2), \alpha_2) = c_0(a_2), \quad \frac{\partial\gamma}{\partial z}(E_0(\alpha_2), \alpha_2) \equiv 0, \\ &\frac{\partial^2\gamma}{\partial z^2}(E_0(\alpha_2), \alpha_2) < 0, \quad \frac{\partial\gamma}{\partial\alpha_2}(0,0) = 0. \end{aligned} \tag{5.4b}$$

Thus, by the Taylor expansion at $z = E_0(\alpha_2)$ we can easily see that $\alpha_1 = \gamma(z, \alpha_2) < c_0(\alpha_2)$ for $z \neq E_0(\alpha_2)$. Indeed, expanding γ at $z = E_0(\alpha_2)$ and taking the square root, we have

$$\pm\sqrt{c_0(\alpha_2) - \alpha_1} = \sqrt{-\tfrac{1}{2}\partial^2\gamma/\partial z^2(E_0(\alpha_2), \alpha_2) + O(|z - E_0(\alpha_2)|)}(z - E_0(\alpha_2)). \tag{5.4c}$$

Therefore, for every $\alpha_1 < c_0(\alpha_2)$ there are exactly two equilibria lying on both sides of $E_0(\alpha_2)$. Denote the one above E_0 by E_+ and the other by E_-. Note that E_+ and E_- collide at E_0 when $\alpha_1 = c_0(\alpha_2)$. As we have mentioned earlier, in the Introduction, a number of people have also contributed to the following theorem.

THEOREM 5.1 (Chow and Lin (1988)). *For a generic two-parameter family of vector fields satisfying conditions* (5.1a, b) *and* (5.2a) *for a nondegenerate homoclinic orbit to a saddle-node equilibrium there exists in a neighborhood* Λ *of* $\alpha = 0$ *a* C^{r-3} *curve*

$\alpha_1 = c_1(\alpha_2)$ with a quadratic tangency to the c_0 curve at $\alpha = 0$ (i.e., $c_0(0) = c_1(0)$, $c_0'(0) = c_1'(0)$ but $c_0''(0) \neq c_1''(0)$) such that, up to possibly renaming the direction of α_2, the following are satisfied in a small tubular neighborhood of the homoclinic orbit:

(i) *For $\alpha \in \mathrm{I} = \{\alpha \in \Lambda \mid$ either $\alpha_1 > c_0(\alpha_2)$, $\alpha_2 \leqq 0$, or $\alpha_2 > 0$, $\alpha_1 > c_1(\alpha_2)\}$ there exists a unique hyperbolic periodic orbit having m Floquet multipliers inside the unit circle in the plane.*

(ii) *For $\alpha \in \mathrm{II} = \{\alpha \in \Lambda \mid \alpha_1 = c_1(\alpha_2), \alpha_2 > 0\}$ there exists a unique homoclinic orbit to E_+.*

(iii) *For $\alpha \in \mathrm{III} = \{\alpha \in \Lambda \mid \alpha_1 < c_1(\alpha_2), \alpha_2 \geqq 0$ or $\alpha_1 < c_0(\alpha_2), \alpha_2 \leqq 0\}$ there exists a unique global heteroclinic orbit from E_+ to E_- in addition to the one connecting E_+ to E_- from the local bifurcation of the saddle-node equilibrium. In particular, when $\alpha_1 < c_1(\alpha_2)$, $\alpha_2 > 0$, respectively, $\alpha_1 = c_1(\alpha_2)$, $\alpha_2 < 0$, respectively, $c_1(\alpha_2) < \alpha_1 \leqq c_0(\alpha_2)$, $\alpha_2 \leqq 0$, this orbit approaches E_- forward in time asymptotically along the center manifold from above E_-, respectively, the strong stable manifold of E_-, respectively, the center manifold from below E_-* (see Fig. 5.2).

Next, we consider the transcritical case (5.3b). To find nonzero equilibrium points we solve equation $\tilde{\theta}(z, \alpha) = 0$ for z, where $\tilde{\theta} = \theta(z, \alpha)/z$. Again by the IFT we obtain a C^{r-2} function $z = E_1(\alpha)$ satisfying

$$E_1(0) = 0, \quad \frac{\partial E_1}{\partial \alpha_1}(0) < 0, \quad \frac{\partial E_1}{\partial \alpha_2}(0) = 0. \tag{5.5}$$

Then we have Theorem 5.2 below.

THEOREM 5.2. *For a generic two-parameter family of vector fields satisfying* (5.1a, b) *and* (5.2b) *for a nondegenerate homoclinic orbit to a transcritical equilibrium there exists in a neighborhood Λ of $\alpha = 0$ a C^{r-2} curve $\alpha_1 = c_1(\alpha_2)$ satisfying $c_1'(0) < 0$ such that, up*

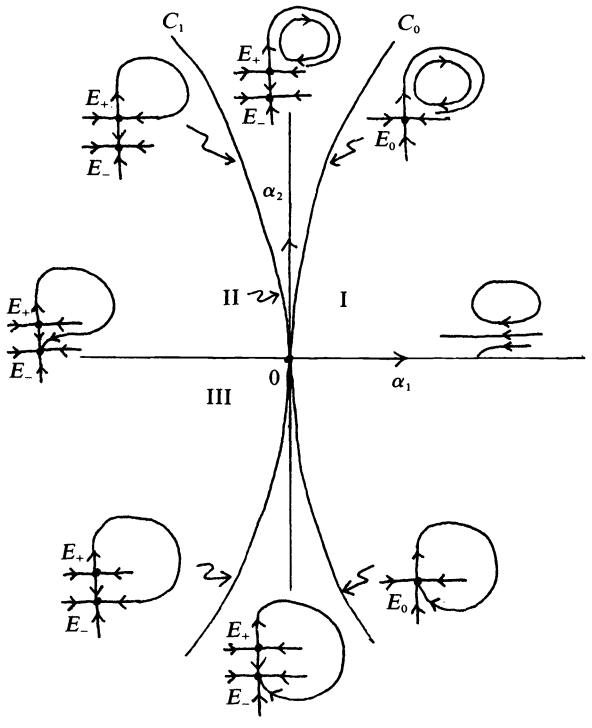


FIG. 5.2. *The bifurcation diagram for the saddle-node homoclinic bifurcation.*

to possibly renaming the direction of α_2, the following are satisfied in a small tubular neighborhood of the homoclinic orbit:

(i) *For $\alpha \in \mathrm{I} = \{\alpha \in \Lambda \mid \alpha_1 > c_1(\alpha_2),\ \alpha_2 > 0\}$ there exists a unique periodic orbit having m Floquet multipliers inside the unit circle.*

(ii) *For $\alpha \in \mathrm{II} = \{\alpha \in \Lambda \mid \alpha_1 = c_1(\alpha_2),\ \alpha_2 > 0\}$ there exists a unique homoclinic orbit to E_1.*

(iii) *For $\alpha \in \mathrm{III} = \{\alpha \in \Lambda \mid \alpha_1 < c_1(\alpha_2)\}$ there exists a unique global heteroclinic orbit from E_1 to the origin in addition to the local connection due to the local transcritical bifurcation. In particular, it approaches the origin from different sides of the origin as the sign of α_2 changes, and from the strong stable manifold of the origin when $\alpha_2 = 0$.*

(iv) *For $\alpha \in \mathrm{IV} = \{\alpha \in \Lambda \mid \alpha_1 = 0,\ \alpha_2 \leqq 0\}$ there exists a unique homoclinic orbit to the saddle-node origin.*

(v) *For $\alpha \in \mathrm{V} = \{\alpha \in \Lambda \mid \alpha_1 > 0,\ \alpha_2 < 0\}$ there exists a unique global heteroclinic orbit from the origin to E_1 in addition to the local connection. In particular, it approaches E_1 from its different sides on the center manifold as α crosses the curve $\alpha_1 = c_1(\alpha_2)$, and from the strong unstable manifold of E_1 on the curve.*

(vi) *For $\alpha \in \mathrm{VI} = \{\alpha \in \Lambda \mid \alpha_2 = 0,\ \alpha_1 > 0\}$ there exists a unique homoclinic orbit to the origin* (see Fig. 5.3).

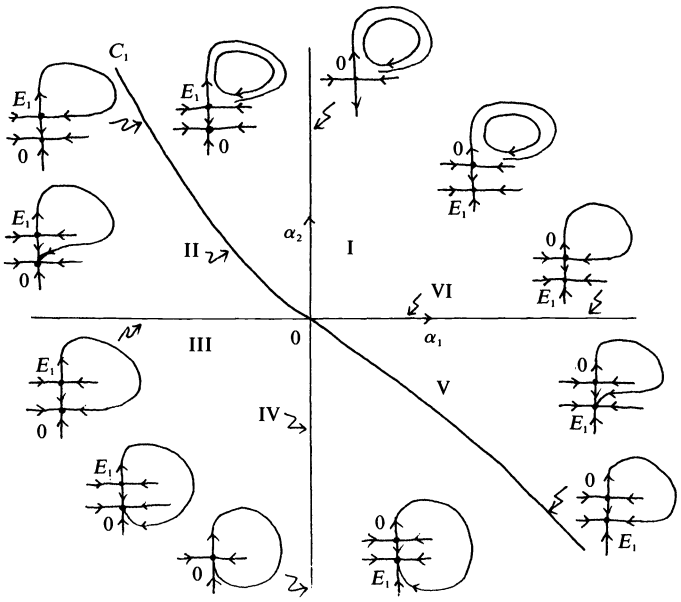


FIG. 5.3. *The bifurcation diagram for the transcritical homoclinic bifurcation.*

Finally, we consider the pitchfork case (5.3c). To find nonzero equilibrium points we solve the equation $\tilde{\theta}(z, \alpha) = 0$ for α_1 by the IFT, where $\tilde{\theta} = \theta(z, \alpha)/z$ and obtain a C^{r-2} curve $\alpha_1 = \gamma(z, \alpha_2)$ satisfying

$$\gamma(0, \alpha_2) = \frac{\partial \gamma}{\partial z}(0, \alpha_2) = 0, \qquad \frac{\partial^2 \gamma}{\partial z^2}(0, 0) < 0. \tag{5.6a}$$

Thus, by the Taylor expansion at $z = 0$, we can easily see that $\alpha_1 = \gamma(z, \alpha_2) < 0$ for $z \neq 0$. In fact, we have

$$\pm\sqrt{-\alpha_1} = \sqrt{-\tfrac{1}{2}(\partial^2 \gamma / \partial z^2)(0, \alpha_2) + O(|z|)}\, z. \tag{5.6b}$$

Therefore, for every $\alpha_1 < 0$ there exist exactly two nonzero equilibria lying on both sides of the zero on the z-axis. Denote the one above the origin by E_+ and the other by E_-. Note that E_+ and E_- collide at the origin when $\alpha_1 = 0$. We have Theorem 5.3.

THEOREM 5.3. *For a generic two-parameter family of vector fields satisfying conditions* (5.1a, b) *and* (5.2c) *for a nondegenerate homoclinic orbit to a pitchfork equilibrium, there exists in a neighborhood* Λ *of* $\alpha = 0$ *a* C^{r-4} *curve* $\alpha_1 = c_1(\alpha_2)$ *with a quadratic tangency to the* α_2*-axis (i.e.,* $c_1(0) = c_1'(0) = 0$ *but* $c_1''(0) \neq 0$*) such that, up to possibly renaming the direction of* α_2*, the following are satisfied in a small tubular neighborhood of the homoclinic orbit:*

(i) *For* $\alpha \in \mathrm{I} = \{\alpha \in \Lambda \mid \alpha_1 > c_1(\alpha_2), \alpha_2 > 0\}$ *there exists a unique periodic orbit having m Floquet multipliers inside the unit circle.*

(ii) *For* $\alpha \in \mathrm{II} = \{\alpha \in \Lambda \mid \alpha_1 = c_1(\alpha_2), \alpha_2 > 0\}$ *there exists a unique homoclinic orbit to* E_+.

(iii) *For* $\alpha \in \mathrm{III} = \{\alpha \in \Lambda \mid \alpha_1 < c_1(\alpha_2)\}$ *there exists a unique global heteroclinic orbit from* E_+ *to the origin, approaching the origin asymptotically along the center direction but from its different sides as the sign of* α_2 *changes.*

(iv) *For* $\alpha \in \mathrm{IV} = \{\alpha \in \Lambda \mid \alpha_1 = c_1(\alpha_2), \alpha_2 < 0\}$ *there exists a unique global heteroclinic orbit from* E_+ *to* E_-.

(v) *For* $\alpha \in \mathrm{V} = \{\alpha \in \Lambda \mid \alpha_1 > c_1(\alpha_2), \alpha_2 < 0\}$ *there does not exist any global homoclinic, heteroclinic, or periodic orbit.*

(vi) *For* $\alpha \in \mathrm{VI} = \{\alpha \in \Lambda \mid \alpha_2 = 0, \alpha_1 > 0\}$ *there exists a unique homoclinic orbit to the origin which is the continuation of the original homoclinic orbit* (see Fig. 5.4).

Before proving the theorems, let us draw heuristically the phase portraits in Fig. 5.5 for some oversimplified situations where $d = 2$, $\theta = \alpha_1 + z^2$ for the saddle-node case, $\theta = z(\alpha_1 + z)$ for the transcritical case, and $\theta = z(\alpha_1 + z^2)$ for the pitchfork case, respectively. In terms of the straight invariant foliation, the c_1 curves, for example, are given

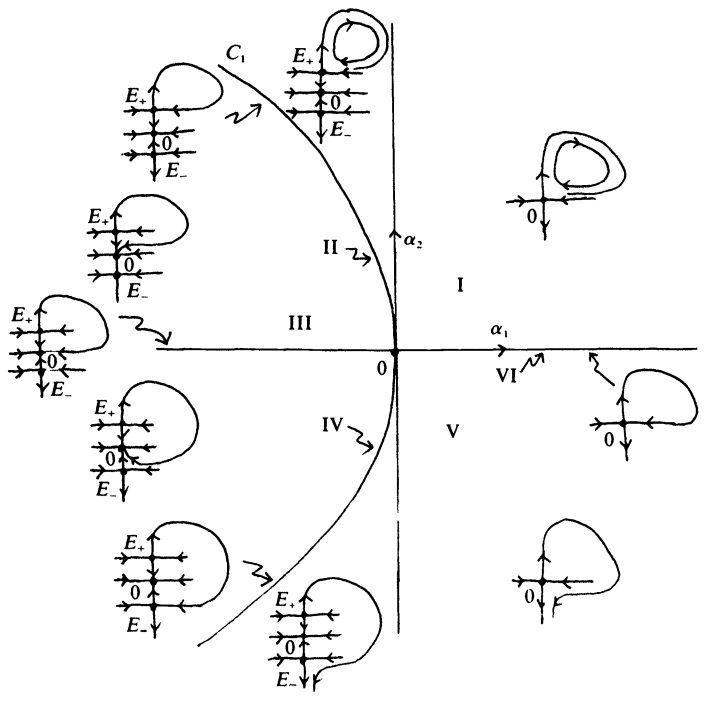


FIG. 5.4. *The bifurcation diagram for the pitchfork homoclinic bifurcation.*

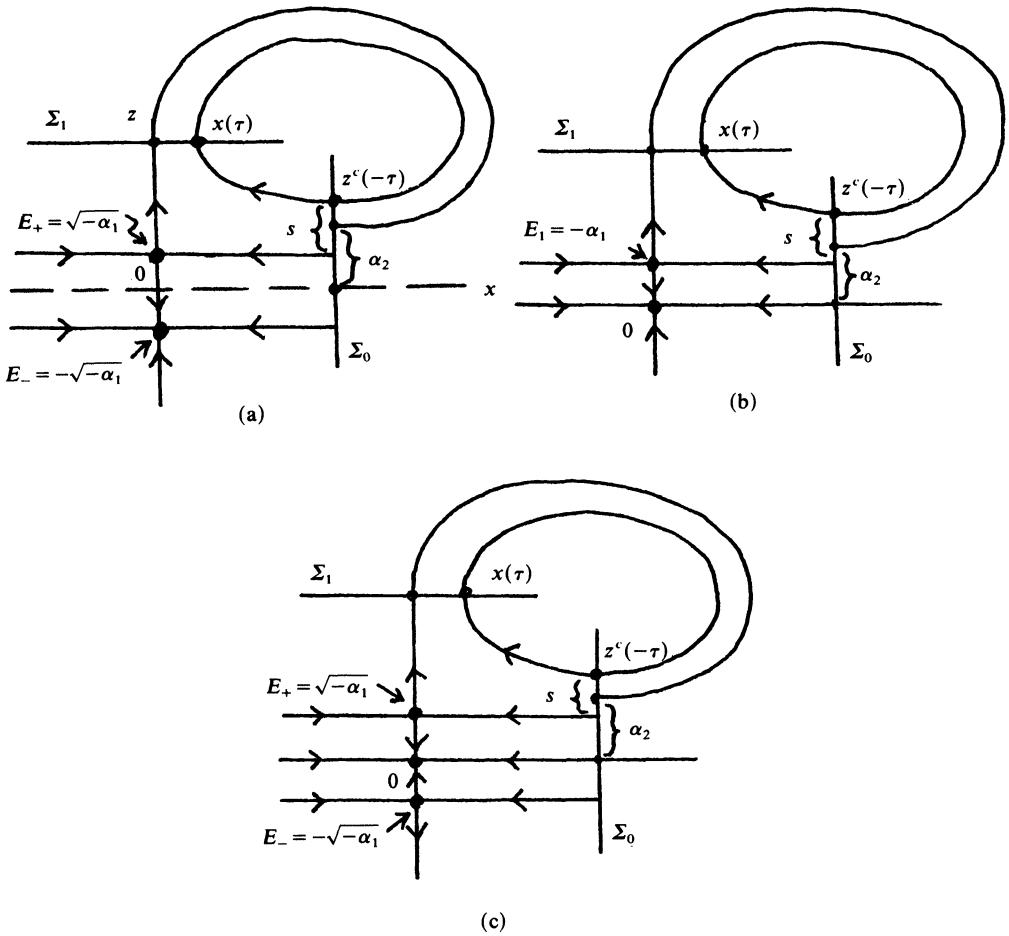


FIG. 5.5. *Some phase portraits for when periodic orbits take place.* (a) *Saddle-node.* (b) *Transcritical.* (c) *Pitchfork.*

as $\alpha_1 = -\alpha_2^2$ in the first and third cases and $\alpha_1 = -\alpha_2$ in the second case. Also, the existence of the periodic orbits is equivalent to solving the scalar equation $z^c(0) = \Pi_1(x(\tau), \delta_0)$ for the time τ, where $(x(t), z^c(t))$ solves the second type of Sil'nikov problem $x(0) = x_0 = \delta_0$ and $z^c(\tau) = z_1 = \delta_0$. Let $E_+(\alpha_1)$ denote the bifurcated equilibrium point above the origin, if any, or zero otherwise, and let $s = z^c(0) - E_+$ be the distance of the "initial" point $z^c(0)$ on Σ_0 to E_+. Then $x(\tau)$ must be $O(s^2)$ by the exponential expansion (for more details concerning this estimate, see the proof below). Thus, we obtain the bifurcation equation

$$s = -E_+(\alpha_1) + \alpha_2 + O(s^2) \quad \text{for } s > 0 \tag{5.7}$$

by the Taylor expansion $\Pi_1(x, \delta_0) = \alpha_2 + O(x)$. We do not even have the constraint $s > 0$ in the saddle-node bifurcation case, where the equilibrium disappears completely for $\alpha_1 > 0$. Using this equation together with the straight foliation, it is not difficult to derive all the conclusions in the theorems. Not surprisingly, we will derive the same form of the bifurcation equation (5.7) for the general cases through a modified Lyapunov-Schmidt reduction. Motivated by these model examples, we can now prove Theorem 5.1.

Proof of Theorem 5.1. We will assume that our readers are familiar with the construction of the Poincaré cross sections Σ_0 and Σ_1 and the Poincaré return maps Π_0 and Π_1 from § 1. The necessary modifications are made as follows: the y-component there is now augmented into the (y, z)-component and, specifically, Σ_1 is given as $\{z = \delta_0\}$ in the δ_0-box of the origin. To use the idea of Lyapunov-Schmidt reduction to the return map $\Pi_1 \circ \rho_1(\zeta) = \rho_0(\zeta)$ under the Sil'nikov variables, we need to normalize our variables in the following as preparations.

First, normalize the local coordinates on Σ_0 as (ξ, y, z) such that $(\xi, y, z) = (0, 0, 0)$ represents the intersection point $\Gamma \cap \Sigma_0$. Similarly, use (x, η) for Σ_1 so that $(x, \eta) = (0, 0)$ corresponds $\Gamma \cap \Sigma_1$ because we have assumed the homoclinic orbit is asymptotically tangent to the center eigenvector space as $t \to -\infty$. See Fig. 5.6.

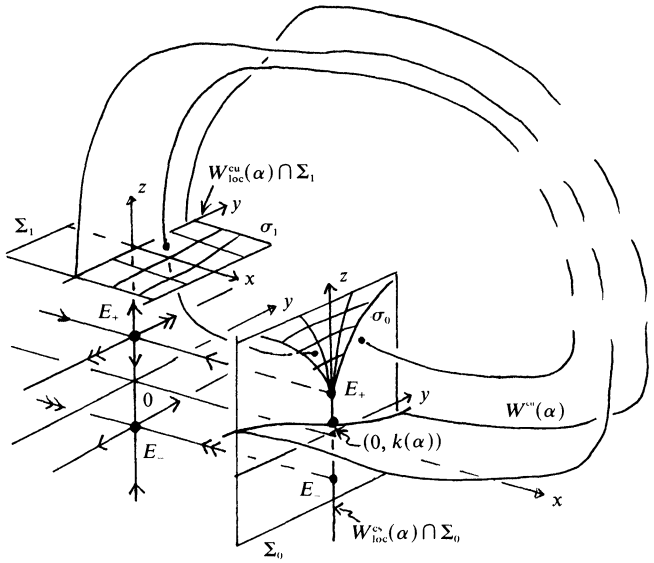


FIG. 5.6. *The cross sections and a perturbed phase portrait for the saddle-node case.*

Second, use Lemma 3.1 for the second type of exponential expansion together with Remark 3.2(a) and expand the Sil'nikov solution with respect to the center trajectory $z^c(t - \tau, \delta_0, \alpha)$ for $0 \leqq t \leqq \tau$ satisfying $z^c(0, \delta_0, \alpha) = \delta_0$. The Sil'nikov variables $(\tau, \xi, \eta, \alpha)$ parametrize Σ_0 and Σ_1 as follows:

$$(\tau, \xi, \eta, \alpha) \to (\xi, y(0, \tau, \xi, \eta, \delta_0, \alpha), z(0, \tau, \xi, \eta, \delta_0, \alpha)),$$

$$(\tau, \xi, \eta, \alpha) \to (x(\tau, \tau, \xi, \eta, \delta_0, \alpha), \eta, \delta_0).$$

Replace τ by a variable s, where τ and s are related by

$$s = z^c(-\tau, \delta_0, \alpha) - s^*(\alpha),$$

where

$$s^*(\alpha) = \begin{cases} E_+(\alpha) & \text{if } \alpha_1 \leqq c_0(\alpha_2), \\ E_0(\alpha_2) & \text{otherwise.} \end{cases}$$

Since $\partial z^c / \partial \tau \neq 0$, we may solve for τ as a function of (s, α):

$$\tau = \tau(s, \alpha). \tag{5.8}$$

Recall from (3.1a) that

$$z(0, \tau, \xi, \eta, \delta_0, \alpha) = z^c(-\tau, \delta_0, \alpha) + R(0, \tau, \xi, \eta, \delta_0, \alpha).$$

Define

$$X(s, \xi, \eta, \alpha) = x(\tau, \tau, \xi, \eta, \delta_0, \alpha),$$
$$Y(s, \xi, \eta, \alpha) = y(0, \tau, \xi, \eta, \delta_0, \alpha),$$
$$R(s, \xi, \eta, \alpha) = R(0, \tau, \xi, \eta, \delta_0, \alpha),$$

where τ is defined by (5.8). Then the normalized Sil'nikov variables (s, ξ, η, α) parametrize Σ_0 and Σ_1 as follows:

$$\rho_0(s, \xi, \eta, \alpha) = (\xi, Y(s, \xi, \eta, \alpha), s + s^*(\alpha) + R(s, \xi, \eta, \alpha)),$$
$$\rho_1(s, \xi, \eta, \alpha) = (X(s, \xi, \eta, \alpha), \eta).$$

Clearly, the local map under these new variables is

$$\Pi_0 \circ \rho_0(s, \xi, \eta, \alpha) = \rho_1(s, \xi, \eta, \alpha).$$

Note that the change of variables $\tau \to s$ is C^{r-1} in τ and at least continuous in α. Actually we will see in a moment that it is C^{r-3} in ε and α_2 if $\alpha_1 \leqq c_0(\alpha_2)$, where $\varepsilon = \sqrt{c_0(\alpha_2) - \alpha_1}$, and C^{r-1} in $\alpha_1 > c_0(\alpha_2)$. Note also that when $\alpha_1 > c_0(\alpha_2)$ there is not any equilibrium point; therefore, the existence time of definition for the local center trajectory $z^c(-\tau, \delta_0, \alpha)$ is finite and s can be both positive and negative depending on whether τ sufficiently large. On the contrary, we require $s > 0$ when $\alpha_1 \leqq c_0(\alpha_2)$.

Third, we also need to extend the functions X, Y, and R differentiably to $s \leqq 0$ whenever $\alpha_1 \geqq c_0(\alpha_2)$ occurs. To do this, it suffices to show that these functions are of order $O(s^2)$, at least at this parameter range. Because of the exponential bounds $e^{\lambda\tau}$ and $e^{-u\tau}$ for the functions $x(\tau, \tau, x_0, y_1, z_1, \alpha)$ and $(y, z)(0, \tau, x_0, y_1, z_1, \alpha)$, respectively, it suffices to show $s = z^c(-\tau, \delta_0, \alpha) - E_+(\alpha) \geqq a\, e^{-b\varepsilon\tau}$ for some positive constants a and b. To show the lower exponentially small bound on $\alpha_1 = c_0(\alpha_2)$, we take the Taylor expansion of θ at $z = E_0(\alpha_2) = E_+(\alpha)$ and use (5.3a) to obtain

$$\theta(z, \alpha) = \frac{1}{2}\frac{\partial^2\theta}{\partial z^2}(E_0, \alpha)(z - E_0)^2 + O(|z - E_0|^3)$$
$$\leqq a_1(z - E_0)^2$$

for some positive constant a_1. Thus, by the comparison principle, the integral curve through the same value δ_0 at $t = 0$ for the vector field $\theta(z, \alpha)$ lies above that of the vector field $a_1(z - E_0)^2$ for the negative time, that is,

$$z^c(-\tau, \delta_0, \alpha) - E_0(\alpha_2) \geqq \frac{k_1}{\tau + k_2} > 0$$

for some constants k_i. This certainly implies the desired lower bound. For the other case where $\alpha_1 < c_0(\alpha_2)$, we take the Taylor expansion of θ at $z = E_+(\alpha)$: $\theta(z, \alpha) = \partial\theta/\partial z(E_+, \alpha)(z - E_+) + O(|z - E_+|^2)$. Expanding $\partial\theta/\partial z$ further, we have

$$\theta(z, \alpha) = \left[\frac{\partial^2\theta}{\partial z^2}(E_0, \alpha)(E_+ - E_0) + O(|E_+ - E_0|^2)\right](z - E_+) + O(|z - E_+|^2)$$
$$\leqq b\varepsilon(z - E_+)$$

for some positive constant b since, by (5.4c), $E_+ - E_0 = O(\sqrt{c_0(\alpha_2) - \alpha_1}) = O(\varepsilon)$. (In fact, by the IFT we can solve $E_+ - E_0$ as a C^{r-3} function of ε and α_2.) Therefore, by the comparison principle we again derive the desired lower bound for s. We also use X, Y, and R to extend the functions.

As the last preparation, we write the global map Π_1 in the normalized coordinates for Σ_0 and Σ_1 as

$$\xi = P(x, \eta, \alpha), \quad y = Q(x, \eta, \alpha), \quad z = T(x, \eta, \alpha).$$

We are now ready to consider the equation $\rho_0(s, \xi, \eta, \alpha) = \Pi_1 \circ \rho_1(s, \xi, \eta, a)$ for periodic orbits running around the homoclinic loop once. This is equivalent to solving the equation $\Phi(s, \xi, \eta, \alpha) = 0$ for the normalized Sil'nikov variable (s, ξ, η) with the constraint $s > 0$ only if $\alpha_1 \leqq c_0(\alpha_2)$, where

$$\Phi(s, \xi, \eta, \alpha) = \begin{pmatrix} \xi \\ Y(s, \xi, \eta, \alpha) \\ s + s^*(\alpha) + R(s, \xi, \eta, \alpha) \end{pmatrix} - \begin{pmatrix} P \\ Q \\ T \end{pmatrix}(X(s, \xi, \eta, \alpha), \eta, \alpha).$$

Certainly $\Phi(0, 0, 0, 0) = 0$ due to the existence of the homoclinic orbit. Compute the Jacobian with respect to (s, ξ, η) at the origin in order to use the IFT; then we have

$$D\Phi(0, 0, 0, 0) = \begin{bmatrix} 0 & I & -P_2'(0, 0, 0) \\ 0 & 0 & -Q_2'(0, 0, 0) \\ 1 & 0 & -T_2'(0, 0, 0) \end{bmatrix} \tag{5.9}$$

by (3.1c), Remark 3.2(b), and the definition of s. Since the first m-columns span the local center-stable tangent space in Σ_0, $T_p(W^{cs}_{loc} \cap \Sigma_0)$, at $p = \Gamma \cap \Sigma_0$ while the last n-columns span the global center-unstable tangent space in Σ_0, $T_p(W^{cu} \cap \Sigma_0)$, then by the nondegeneracy condition (5.1a) the Jacobian is nonsingular. Thus, by the IFT, a unique solution $(s, \xi, \eta) = (\bar{s}, \bar{\xi}, \bar{\eta})(\alpha)$ exists for α from a small neighborhood Λ of the origin. To ensure that this solution indeed gives rise to a periodic orbit, we need to find only those α satisfying $\alpha_1 \leqq c_0(\alpha_2)$ such that the constraint $s > 0$ is satisfied because there is no restriction on s when $\alpha_1 > c_0(\alpha_2)$. To do this, we need the following Lyapunov-Schmidt reduction procedure to obtain the bifurcation equation.

Because of the special structure of the Jacobian (5.9) of Φ, we can first solve ξ and η in terms of s and α from the first $m + n - 1$ equations of $\Phi = 0$ by the IFT. Thus, as far as only these equations are concerned, $\bar{\xi}(\alpha)$ and $\bar{\eta}(\alpha)$ are the solutions and $\bar{s}(\alpha)$ can be treated as an independent variable. Hence, by formally setting $\bar{s} = 0$ and plugging $\xi = \bar{\xi}(\alpha)$ and $\eta = \bar{\eta}(\alpha)$ in the function Φ, the equation $\Phi = 0$ will be respected except for the last equation, that is,

$$\bar{\xi} = P(0, \bar{\eta}, \alpha), \quad 0 = Q(0, \bar{\eta}, \alpha), \quad k(\alpha) = T(0, \bar{\eta}, \alpha),$$

where the function $k(\alpha) \overset{\text{def}}{=} T(0, \bar{\eta}, \alpha)$. Note that the geometrical interpretation of this relation is that $(\bar{\xi}, 0, k(\alpha))$ is the unique intersection point of the global center-unstable manifold $\{(P, Q, T)(0, \eta, \alpha)\}$ with the local center-stable manifold $W^{cs}_{loc} = \{y = 0\}$ (cf. Fig. 5.6) in Σ_0. By the transversality condition (5.1a) the function $k(\alpha)$ must be at least C^{r-2} (the same as the admissible variable; the same conclusion also holds for $\bar{\xi}(\alpha)$ and $\bar{\eta}(\alpha)$ as well). By means of the distance $d(\alpha)$ between W^{cu} and W^{s}_{loc} in Σ_0, we actually know more about the function $k(\alpha)$. Indeed, by definition, it must satisfy

$$0 \leqq d(\alpha) = \min_{|\xi|, |\eta| \leqq \delta_0} |(P, Q, T)(0, \eta, \alpha) - (\xi, 0, 0)| \leqq |k(\alpha)|,$$

$$\lim_{\alpha_2 \to 0} \frac{|k((0, \alpha_2))|}{|\alpha_2|} > 0$$

by the transverse crossing condition (5.1b). Thus, for $\alpha_2 > 0$, $k((0, \alpha_2))$ must have a constant sign. Since $k(0) = 0$, the inequalities above imply

$$\frac{\partial k(0)}{\partial \alpha_2} \neq 0.$$

For definiteness, we assume $\partial k(0)/\partial \alpha_2 > 0$, which corresponds to preserving the direction of α_2 in the statement of the theorem.

Now, the desired bifurcation equation is simply the last equation of $\Phi = 0$ at $(s, \xi, \eta) = (\bar{s}, \bar{\xi}, \bar{\eta})(\alpha)$ and $\alpha_1 \leqq c_0(\alpha_2)$. Using Taylor expansion at $\bar{s} = 0$, and the order estimates for the functions X, Y, and R above, we have

$$\bar{s} = -E_+(\alpha) + k(\alpha) + O(\bar{s}^2),$$

which has the same form as (5.7). Thus $\bar{s} > 0$ for $\alpha_1 \leqq c_0(\alpha_2)$ if and only if

$$k(\alpha) > E_+(\alpha).$$

To describe this region $\{k(\alpha) > E_+(\alpha)\} \cap \{\alpha_1 \leqq c_0(\alpha_2)\}$, let us begin with its boundary $k(\alpha) = E_+(\alpha)$. This is precisely when the homoclinic orbit to $E_+(\alpha)$ takes place, since the stable manifold of E_+ is $\{y = 0, z = E_+\}$ by the "straight" foliation of the admissible variable mentioned in §§ 1 and 2 and the intersection point of the unstable manifold of E_+, where the center-stable manifold of the origin is $(\bar{\xi}, 0, k(\alpha))$. By substituting $z = E_+(\alpha) = k(\alpha)$ into the function $\alpha_1 = \gamma(z, \alpha_2)$ in (5.4b) we can solve for a C^{r-2} curve $\alpha_1 = c_1(\alpha_2)$ satisfying $c_1'(0) = 0$ from the equation $\alpha_1 = \gamma(k(\alpha), \alpha_2)$, and $c_1(\alpha_2) < c_0(\alpha_2)$ is always true by the definition of γ. Thus $k(\alpha) = E_+(\alpha)$, or $E_-(\alpha)$ on c_1. Indeed we claim that $k(\alpha) = E_+(\alpha)$ is satisfied if and only if $\alpha_1 = c_1(\alpha_2)$ and $\alpha_2 > 0$. To show this we need only to rule out $k(\alpha) = E_+(\alpha)$ on the lower half c_1 curve. Since $E_0'(0) = 0$, $c_1'(0) = 0$ and $d/d\alpha_2 k((c_1, \alpha_2))|_{\alpha_2 = 0} = \partial k/\partial \alpha_2(0) > 0$, $k((c_1(\alpha_2), \alpha_2)) = E_+((c_1(\alpha_2), \alpha_2)) \geqq E_0(\alpha_2)$ if and only if $\alpha_2 > 0$. For exactly the same reason we see that $k((c_0(\alpha_2), \alpha_2)) > E_+((c_0(\alpha_2), \alpha_2)) = E_0(\alpha_2)$ for $\alpha_2 > 0$ and $k((c_0(\alpha_2), \alpha_2)) < E_-((c_0(\alpha_2), \alpha_2)) = E_0(\alpha_2)$ for $\alpha_2 < 0$. Moreover, since $E_+(\alpha) - E_0(\alpha_2) = O(\sqrt{c_0(\alpha_2) - \alpha_1})$ while $k(\alpha) - E_0(\alpha_2)$ is differentiable, it must be that $k(\alpha) < E_+(\alpha)$ for $\alpha_1 < c_1(\alpha_2)$. Therefore, we can conclude that $\{k(\alpha) > E_+(\alpha)\} \cap \{\alpha_1 \leqq c_0(\alpha_2)\}$ is the wedge-shaped region between the curves c_1 and c_0:

$$\{\alpha \mid c_1(\alpha_2) < \alpha_1 \leqq c_0(\alpha_2), \alpha_2 > 0\}$$

(see Fig. 5.7).

To show that c_1 has quadratic tangency to the c_0 curve, simply observe from (5.4c), $c_1'(0) = c_0'(0) = E_0'(0) = 0$, and $\partial k/\partial \alpha_2(0) \neq 0$ that

$$(c_1 - c_0)''(0) = \frac{1}{2} \frac{\partial^2 \gamma}{\partial z^2}(0, 0) \left[\frac{\partial k}{\partial \alpha_2}(0) \right]^2 < 0.$$

Next, to show that the periodic orbit is unique, we need to rule out the possibility that there might be solutions to the following cyclic equations other than the trivial one $(\bar{s}, \bar{\xi}, \bar{\eta})$ found above:

$$\begin{pmatrix} \xi_{i+1} \\ Y(s_{i+1}, \xi_{i+1}, \eta_{i+1}, \alpha) \\ s_{i+1} + s^*(\alpha) + R(s_{i+1}, \xi_{i+1}, \eta_{i+1}, \alpha) \end{pmatrix} - \begin{pmatrix} P \\ Q \\ T \end{pmatrix} (X(s_i, \xi_i, \eta_i, \alpha), \eta, \alpha) = 0$$

for $i = 0, \cdots, (\mathrm{mod}\ k)$. Note that their solutions with $s_i > 0$ imply the existence of periodic orbits running around the loop k times. By the IFT again, we can show that

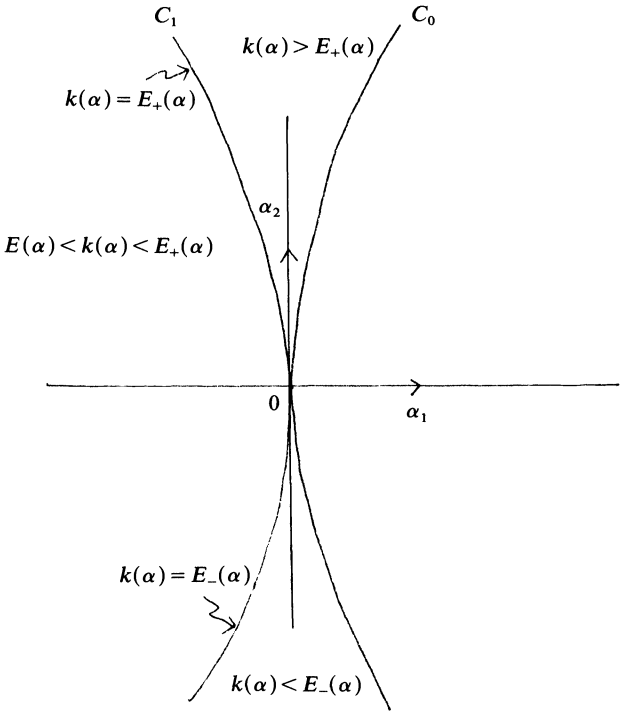


FIG. 5.7

the solutions are unique, which must be the repetition of k copies of the $(\bar{s}, \bar{\xi}, \bar{\eta})$. Thus the uniqueness is established. Indeed, because the associated Jacobian consists of nonzero blocks only as does $D\Phi(0, 0, 0, 0)$ above, the parameter range on which the existence and uniqueness conclusion holds for such periodic orbits can be chosen the same as Λ.

Thus to complete (i) it only remains to show that the periodic orbit has m Floquet multipliers inside and n outside the unit circle. To see this, consider the characteristic polynomial $\det|D(\rho_0^{-1} \circ \Pi \circ \rho_0(\bar{s}, \bar{\xi}, \bar{\eta}, \alpha)) - \lambda I| = 0$, where $\Pi = \Pi_1 \circ \rho_1 \circ \rho_0^{-1}$ is the Poincaré map in the old variables. It is equivalent to considering $\det|D\Pi_1 \cdot D\rho_1(\bar{s}, \bar{\xi}, \bar{\eta}, \alpha) - \lambda D\rho_0(\bar{s}, \bar{\xi}, \bar{\eta}, \alpha)| = 0$, which has the form $\det Q_2'(0, 0, \alpha)\lambda^m + p(\lambda, \bar{s}, \bar{\xi}, \bar{\eta}, \alpha) = 0$, where p is an $(m+n)$-degree polynomial with all the coefficients having the order at least $O(s)$. Thus as $s \to 0$ it has precisely m roots inside and n outside the unit circle.

Part (ii) has been proved above, since $k(\alpha) = E_+(\alpha)$ takes place exactly on the c_1 curve if and only if $\alpha_2 > 0$. To show (iii), note that when $k(\alpha) < E_+(\alpha)$, the point $(\bar{\xi}, 0, k)$ from the global unstable manifold of E_+ lies in the local stable manifold of E_-, which is $\{y = 0\}$. Note also that when $k(\alpha) = E_-(\alpha)$ the heteroclinic connection comes in along the strong stable manifold of E_-, which is $\{y = 0, z = E_-\}$ by the straight foliation of the admissible variable. This happens precisely on the curve $\alpha_1 = c_1(\alpha_2)$ for the same reasons as for $k(\alpha) = E_+(\alpha)$ above. But this time $\alpha_2 < 0$ since $k(\alpha) = E_-(\alpha) < E_+(\alpha)$. □

The proofs for Theorems 5.2 and 5.3 follow the same strategy as above. That is, use s, the distance of the center trajectory $z^c(-\tau, \delta_0, \alpha)$ to the bifurcated equilibrium above the origin; use the comparison principle to estimate the lower bounds of s in terms of an exponentially small number $e^{-\varepsilon\tau}$; extend the functions X, Y, and Z to $s \leqq 0$ differentiably; and use the IFT to obtain the bifurcation equation and the straight

foliation of the admissible variable to establish the connections. In these two cases, s is always positive because of the persistence of equilibria. We omit the details here because the proofs are not only similar but also much easier.

Acknowledgment. I acknowledge my postdoctoral fellowship at the Lefschetz Center for Dynamical Systems and the Program for Turbulence at Brown University from 1987 to 1988. Special thanks go to Professors J. K. Hale, J. Mallet-Paret, and L. Sirovich for providing me with a stimulating and carefree research environment.

REFERENCES

S. N. CHOW, B. DENG, AND B. FIEDLER (1988), *Homoclinic bifurcations at resonant eigenvalues*, J. DDE, April 1990, to appear.

S. N. CHOW AND X.-B. LIN (1988), *Bifurcation of a homoclinic orbit with a saddle-node equilibrium*, preprint.

S. N. CHOW AND K. LU (1988), *C^k center-unstable manifolds*, Proc. Roy. Soc. Edinburgh, 108A, 303-320.

S. N. CHOW, X.-B. LIN, AND K. LU (1988), *Smoothness of invariant foliation in infinite dimensional spaces*, preprint.

S. N. CHOW, B. DENG, AND D. TERMAN (1987), *The bifurcation of a homoclinic and periodic orbit from two heteroclinic orbits*, SIAM J. Appl. Math., 21 (1990), to appear.

P. COLLET AND J.-P. ECKMANN (1980), *Iterated Maps of the Interval as Dynamical Systems*, Birkhäuser, Boston.

B. DENG (1988a), *The Sil'nikov problem, exponential expansion, strong λ-lemma, C^1-linearization and homoclinic bifurcation*, J. Differential Equations, 79 (1989), pp. 189-231.

——— (1988b), *Exponential expansion with Sil'nikov's saddle-focus*, J. Differential Equations, 81 (1989).

——— (1988c), *Sil'nikov problem, invariant manifolds and λ-lemma*, preprint.

——— (1988d), *Exponential expansion with principal eigenvalues*, Proc. Roy. Soc. Edinburgh Sect. A, to appear.

——— (1988e), *The bifurcation of countable connections from a twisted heteroclinic loop*, SIAM J. Math. Anal., submitted.

G. DANGELMAYR, D. ARMBRUSTER, AND N. NEVELING (1985), *A codimension three bifurcation for the laser with saturable absorber*, Z. Phys. B, 59, pp. 365-370.

N. FENICHEL (1979), *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31, pp. 53-98.

J. GUCKENHEIMER AND P. HOLMES (1983), *Nonlinear Oscillations, Dyamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, Berlin, New York.

J. K. HALE (1978), *Ordinary Differential Equations*, Second edition, Krieger, Melbourne, FL.

D. HENRY (1981), *Geometric Theory of Parabolic Equations*, Lecture Notes in Math 840, Springer-Verlag, Berlin, New York.

M. HIRSCH, C. PUGH, AND M. SHUB (1977), *Invariant Manifolds*, Lecture Notes in Math. 583, Springer-Verlag, Berlin, New York.

M. C. IRWIN (1980), *Smooth Dynamical Systems*, Academic Press, New York.

H. K. JU (1988), *Bifurcation of symmetric planar vector fields*, Ph.D. thesis, Michigan State University, East Lansing, MI.

P. LANCASTER (1969), *Theory of Matrices*, Academic Press, New York.

V. I. LUK'YANOV (1982), *Bifurcations of dynamical systems with a saddle-point-separatrix loop*, Differentsial'nye Uravneniya, 18, pp. 1493-1506. (In Russian.) J. Differential Equations, 18, pp. 1049-1059. (In English.)

J. MOSER (1973), *Stable and random motions in dynamical systems*, Ann. of Math. Stud., 77, Princeton University Press, Princeton, NJ.

I. M. OVSYANNIKOV AND L. P. SIL'NIKOV (1986), *On systems with saddle-focus homoclinic curve*, Mat. Sb., 172 (1986). (In Russian.) Math. USSR-Sb., 58 (1987), pp. 557-574. (In English.)

S. SCHECTER (1987), *The saddle-node separatrix-loop bifurcation*, SIAM J. Math. Anal., 18, pp. 1142-1157.

L. P. SIL'NIKOV (1967), *On a Poincaré-Birkhoff problem*, Math. USSR-Sb., 3, pp. 353-371.

——— (1970), *A contribution to the problem of the structure of an extended neighborhood of a rough state of saddle-focus type*, Math. USSR-Sb., 10, pp. 91-102.

J. SOTOMAYOR (1974), *Generic one-parameter families of vector fields*, Inst. Hautes Etudes Sci. Publ. Math., 43, pp. 5-46.

A. VANDERBAUWHEDE AND S. VAN GILS (1987), *Center manifold and contractions on a scale of Banach spaces*, J. Funct. Anal., 72, pp. 209-224.

J. C. WELLS (1976), *Invariant manifolds of nonlinear operators*, Pacific J. Math., 625, pp. 285-293.

E. YANAGIDA (1987), *Branching of double pulse solutions from single pulse solutions in nerve axon equation*, J. Differential Equations, 66, pp. 243-262.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 721-728, May 1990
© 1990 Society for Industrial and Applied Mathematics
009

EXISTENCE, BIFURCATION, AND LIMIT OF SOLUTIONS OF THE SIMILARITY EQUATIONS FOR FLOATING RECTANGULAR CAVITIES AND DISKS*

CHUNQING LU†

Abstract. The differential equation $f''' + Q[Aff'' - (f')^2] = \beta$, ($A \in [0, \infty)$, $Q > 0$, β real, $' = d/dx$ for $x \in [0, 1]$) with the boundary conditions $f(0) = f(1) = f''(1) = f''(0) + 1 = 0$ is considered. Existence of solutions of this boundary value problem is proved for $\beta \leqq 0$. Bifurcation of the number of solutions as A varies is studied, and the limit of convex solutions as $Q \to \infty$ is given.

Key words. similarity equations, existence, bifurcation, limit

AMS(MOS) subject classifications. 34, 76

Introduction. The third-order nonlinear ordinary differential equation

$$f''' + Q[Aff'' - (f')^2] = \beta, \tag{1}$$

where $Q > 0$, $A \in [0, \infty)$, and β are constants, $f = f(x)$ is an unknown function defined in $[0, 1]$, and $' = d/dx$ subject to the boundary conditions

$$f(0) = f(1) = f''(1) = f''(0) + 1 = 0 \tag{2}$$

arises from a reduction by similarity of boundary layer formulation of the Navier-Stokes equations for the distributions of velocity in a low Prandtl number fluid zone in the shape of either a floating rectangular slot or a floating circular disk [5], [6]. The flow in low Prandtl number fluid (liquid metal or silicon) is contained by the lateral solid surfaces and surface tension. The floating zones are assumed to be in a microgravity environment so that the force of gravity is neglected. Existence of solutions of (1)-(2) has been proved in [1] for the following cases:

(1) For given $A > 0$ and for $\beta \in (0, 1]$ there exists at least one $Q > 0$ such that the convex solutions exist.

(2) Let $A = 1$ or 2. Then for given $Q > 0$ there exists at least one β such that the convex solutions exist.

In this paper we continue our study of the existence of solutions for $\beta \leqq 0$, some bifurcation phenomena of the number of solutions as parameter A varies, and the limits of such solutions as $Q \to \infty$. The paper is divided into four sections. In § 1 we prove the existence of solutions in the case $\beta = 0$. The existence of convex solutions for all $A \in [1, 2]$ is given in § 2. In § 3 we study bifurcation phenomena, i.e., we prove that if $A = 1$, then the solution is unique for given Q, and that if $A = 2$, then there exist multiple solutions for some Q. Finally, we prove the existence of solutions for $\beta < 0$, and discuss the limit of the convex solutions as $Q \to +\infty$. Considering $Q = c_1 \operatorname{Re}$ and $\beta = c_2 \cdot (\partial P/\partial x)$ for the floating slot ($A = 1$), and $\beta = c_3 \cdot (\partial P/\partial r)$ for the floating disk ($A = 2$), where c_i are constants, P is the pressure, and Re is the Reynolds number (see [5], [6], and [1]), we see that our results imply that

1) If the pressure P is constant ($\beta = 0$), then there always exist two-cell flows;

2) There exist two-cell and three-cell flows for the floating slot;

3) As the Reynolds number Re becomes large, Q is very large. In this case flows exist with $\beta < 0$; namely, the pressure is monotonously decreasing with respect to x

* Received by the editors December 19, 1988; accepted for publication (in revised form) June 11, 1989.

† Institute of Software, Academia Sinica, Beijing, People's Republic of China. Present address, Department of Mathematics, State University of New York, Buffalo, New York 14214.

for the slot or to r for the disk (see [1]). And as $Q \to \infty$ the velocities of the two-cell flows tend to zero.

For convenience, we introduce the change of variable: $y = 1 - x$. Then the equation takes the following form:

$$g''' - Q[Agg'' - (g')^2] = -\beta \qquad (0 \leqq y \leqq 1) \tag{3}$$

subject to the boundary conditions

$$g(0) = g''(0) = g(1) = g''(1) + 1 = 0. \tag{4}$$

1. Existence of solutions in the case $\beta = 0$.

THEOREM 1. *Given $A \geqq 0$, there exists at least one $Q \geqq 0$ such that there exists at least one solution of the boundary value problem* (3)-(4).

Proof. Set $\beta = 0$ in (3), and make further changes of variables: $\eta = cy$, $\varphi(\eta) = g(\eta/c)$, Q/c, where η is the new variable, and c is a parameter that will be determined in the rest of our proof. Then the problem becomes to prove that there exist numbers $c > 0$ and $Q > 0$ such that the equation

$$\varphi''' - [A\varphi\varphi'' - (\varphi')^2] = 0, \tag{1.1}$$

has solutions satisfying

$$\varphi(0) = \varphi''(0) = 0, \tag{1.2}$$

and

$$\varphi(c) = 0, \qquad \varphi''(c) = -\frac{Q}{c^3}. \tag{1.3}$$

Suppose that $\varphi(\eta)$ is a solution of the initial value problem (1.1) with the initial conditions

$$\varphi(0) = 0, \quad \varphi'(0) = \alpha, \quad \varphi''(0) = 0. \tag{1.4}$$

Our goal is to find an appropriate value of α such that the solution φ of (1.1)-(1.4) across the positive half η-axis at $\eta = c$ with $\varphi''(c)$ is less than zero. We divide the proof of Theorem 1 into two lemmas.

LEMMA 1. *Let $\varphi(\eta)$ be a nontrivial solution of the initial value problem* (1.1)-(1.4). *Then $\alpha \neq 0$ and $\varphi''(\eta) < 0$ for all $\eta > 0$.*

Proof. Since $\alpha = 0$ gives only a trivial solution, $\alpha \neq 0$. Hence, $\varphi'''(0) = -\alpha^2 < 0$ and then $\varphi(\eta) > 0$, $\varphi'''(\eta) < 0$ and $\varphi''(\eta) < 0$ initially. Assume that there exists the first positive point $\eta = x_0$ at which $\varphi'' = 0$. Then $\varphi'''(x_0) = -\varphi'(x_0)^2 \leqq 0$. If $\varphi'(x_0) \neq 0$, then there must be a point $x_1 < x_0$ such that $\varphi''(x_1) > 0$ and a point $x_2 < x_1$ such that $\varphi''(x_2) = 0$. This contradicts the assumption that x_0 is the first such point. If $\varphi'(x_0) = 0$, then $\varphi \equiv 0$, which contradicts the fact that $\alpha \neq 0$. Therefore, $\varphi'' < 0$ for all $\eta > 0$. This completes the proof of Lemma 1.

LEMMA 2. *Let $\varphi(\eta)$ solve* (1.1)-(1.4) *with $\alpha > 0$. Then φ must reach zero somewhere in $(0, \infty)$.*

Proof. We observe that $\varphi(\eta) > 0$ for sufficiently small $\eta > 0$. Hence, $\varphi''' < 0$ as long as $\varphi > 0$ by Lemma 1. The Taylor expansion of φ at $\eta = \eta_0$, where $\eta_0 > 0$ is selected so that $\varphi(\eta_0) > 0$, $\varphi'(\eta_0) > 0$, but $\varphi''(\eta_0) < 0$, gives that $\varphi(\eta) < \varphi(\eta_0) + \varphi'(\eta_0)(\eta - \eta_0) + \varphi''(\eta_0)(\eta - \eta_0)^2/2$ as long as φ remains positive. Thus φ must equal zero somewhere in $(0, \infty)$. Lemma 2 is proved.

Combining Lemmas 1 and 2, we conclude that there exist solutions of (1.1) satisfying (1.2) and (1.3) at some $c > 0$. Consequently, $Q = \varphi''(c)c^3 > 0$. The proof of Theorem 1 is completed.

Remarks. Observe from the proof that if φ is a solution of the boundary value problem, then $\varphi'(0)$ must be positive due to $\varphi'' < 0$. This theorem gives the concave down solutions representing two-cell flows. Also, we see that for any φ with $\varphi'(0) > 0$, solving (1.1)-(1.4) must give a number c (depending on α) at which φ crosses the η-axis, and then a positive number Q. Although we have not proved the uniqueness of such Q, it does not simply imply that different α's will present different Q's because $Q = -\varphi''(c)c^3$. Numerically it was found that the number of Q's corresponding to $\beta = 0$ is 1 (see [1]).

2. Existence of solutions in the case $1 \leqq A \leqq 2$. Since we have proved the existence of solutions for $A = 1$ and 2 in [1], we only consider the case $1 < A < 2$. Applying the transformation used in § 1 with $c = 1$, we get the equation

$$\varphi''' - [A\varphi\varphi'' - \varphi'^2] = -\beta Q, \tag{2.1}$$

together with the boundary conditions

$$\varphi(0) = \varphi''(0) = 0, \tag{2.2}$$

$$\varphi(1) = 0, \qquad \varphi''(1) = -Q. \tag{2.3}$$

THEOREM 2. *For given $Q > 0$ and for given $A \in (1, 2)$ there exist at least one number β and a convex function solving* (2.1)-(2.3).

Proof. Differentiate (2.1). Then

$$\varphi'''' = A\varphi\varphi''' + (A-2)\varphi'\varphi''. \tag{2.4}$$

Let $\varphi'(0) = \lambda$ and $\varphi'''(0) = \mu$. We will apply the shooting argument with the shooting parameters λ and μ to the initial value problem (2.2)-(2.4) with $\varphi'(0) = \lambda$ and $\varphi'''(0) = \mu$. First we need the following two lemmas.

LEMMA 3. *Suppose that φ solves* (2.2)-(2.4) *with $\lambda > 0$ and $\mu < 0$. Then $\varphi''' < 0$ for all $\eta > 0$.*

Proof. Initially, φ and φ' are positive, and φ'' and φ''' are negative. Differentiating (2.4) once again, we obtain

$$\varphi^{(5)} = A\varphi\varphi'''' + 2(A-1)\varphi'\varphi''' - (2-A)\varphi''^2. \tag{2.5}$$

Then, by (2.5), $\varphi'''' < 0$ and then $\varphi'' < 0$ as long as $\varphi' > 0$ and $\varphi''' < 0$. Therefore, φ' becomes zero before φ''' does. Assume $\varphi'(\eta_1) = 0$. Then φ'', φ''', and φ' are less than zero and then $\varphi'''' < 0$ for η greater than and closer to η_1. Since $\varphi'''' < 0$ wherever $\varphi''' = 0$ for $\eta > \eta_1$, $\varphi''' \leqq 0$ for $\eta > \eta_1$. Assume $\varphi'''(\eta_2) = 0$ for some first $\eta_2 > \eta_1$. Then, $\varphi''''(\eta_2) = (A-2)\varphi'(\eta_2)\varphi''(\eta_2) < 0$, which implies $\varphi''' > 0$ for some $\eta < \eta_2$, which contradicts $\varphi'''(\eta_2) < 0$. Lemma 3 now follows.

From the proof of Lemma 3 we see that we have proved the following corollary.

COROLLARY. If φ solves (2.2)-(2.4) with $\lambda > 0$ and $\mu < 0$, then $\varphi'''' < 0$ as long as $\varphi \geqq 0$.

Next we define four sets on the quadrant $\Sigma = \{(\lambda, \mu) | \lambda > 0 \text{ and } \mu < 0\}$ of the real $\lambda - \mu$ plane as follows:

$$\Omega_1 = \{(\lambda, \mu) | \varphi(1) > 0\}, \qquad \Omega_2 = \{(\lambda, \mu) | \varphi(1) < 0\},$$

$$\Omega_3 = \{(\lambda, \mu) | \varphi''(1) > -Q\}, \qquad \Omega_4 = \{(\lambda, \mu) | \varphi''(1) < -Q\}.$$

Observe that these four sets are open and $\Omega_1 \cap \Omega_2$ and $\Omega_3 \cap \Omega_4$ are both empty.

Next, we describe the four sets Ω_i in more detail.

LEMMA 4. 1) *The subset $\{(\lambda, \mu) | \mu < -Q\}$ of Σ lies entirely in Ω_4.*

2) *For each $\mu < 0$ there is a $\lambda_0(\mu)$ such that the region $0 < \lambda < \lambda_0(\mu)$ lies entirely in Ω_2.*

3) *For each* $\lambda \geqq 0$ *there is a* $\mu_0(\lambda) < 0$ *such that the region* $\mu_0(\lambda) < \mu < 0$ *lies entirely in* Ω_3, *and that* $\Omega_2 \cap \Omega_4$ *and* $\Omega_2 \cap \Omega_3$ *are not empty.*

4) *The subset* $\{(\lambda, \mu) | \lambda > (Q/2)\}$ *of* Σ *lies entirely in* $\Omega_1 \cup \Omega_4$.

Proof. Conclusion 1 follows directly from Lemma 3. If $\lambda = 0$ and $\mu < 0$ then $\varphi''' < 0$ for all $\eta > 0$ by (2.4), and hence $\varphi(1) < 0$. By the continuous dependence of solutions on λ we get 2. Similarly, we prove conclusion 3. Note that the conditions $\lambda = 0$ and $\mu = 0$ give the trivial solution. Thus, for given $Q > 0$ there is a $\mu_0(0) < 0$ such that $\varphi''(1) > -Q$, and then $\Omega_2 \cap \Omega_3$ is not empty (Fig. 1). Finally, if $\lambda > Q/2$ and $(\lambda, \mu) \notin \Omega_4$, i.e., $\varphi''(1) > -Q$, then $\varphi(1) > \lambda + \varphi''(1)/2 > 0$, so $(\lambda, \mu) \in \Omega_1$. The proof of Lemma 4 is complete.

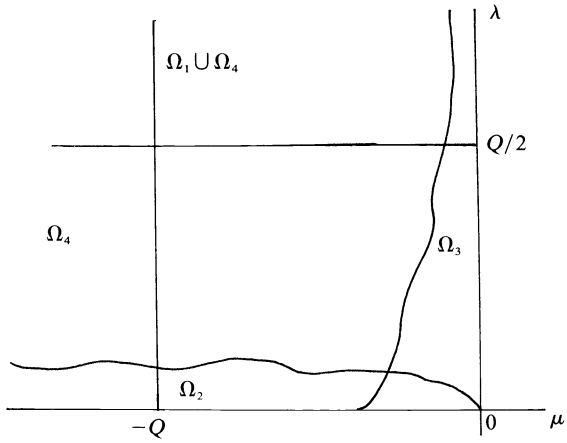


FIG. 1

Lemma 4 is illustrated in Fig. 1. By the lemma of Hastings (see [2] and [3]), it is concluded that the conditions we have found for the four sets show there is a point that is not in $\Omega = \Omega_1 \cup \Omega_2 \cup \Omega_3 \cup \Omega_4$. This proves Theorem 2.

3. Bifurcation phenomena on numbers of solutions. Let $Q > 0$ be given in this section. We will study the number of solutions of (3)-(4) for $A = 1$ and $A = 2$, because most similarity equations in studying flows for the floating zones are concerned with $A = 1$ or 2 (see [5], [6], and [1]). This also implies that the number of β for given Q bifurcates as A varies from 1 to 2. In this section we give a proof of the bifurcation diagram shown analytically in [1].

THEOREM 3. *Suppose* $A = 2$. *Given* $Q \geqq 0$, *there exists a unique number* β *such that the boundary value problem* (3)-(4) *has a unique solution.*

Proof. It is sufficient to prove that for given $Q > 0$ the fourth-order nonlinear equation

$$f'''' - 2Qff''' = 0 \tag{3.1}$$

with boundary conditions (2) has a unique solution. Suppose, by contradiction, that both f_1 and f_2 solve (2.1)-(2.4), and $f_1 \neq f_2$. Define a function $h(\eta) = f_1(\eta) - f_2(\eta)$. Then $h(\eta)$ solves the equation

$$y'''' - 2Qf_1(\eta)y''' - 2Qf_2'''(\eta)y = 0 \qquad (0 \leqq \eta \leqq 1) \tag{3.2}$$

together with the homogeneous boundary conditions:

$$h(0) = h''(0) = 0, \tag{3.3}$$

$$h(1) = h''(1) = 0. \tag{3.4}$$

Without loss of generality, suppose $h'(0) = \lambda \geqq 0$. Recall that $f_i''' < 0$ (see [1]), $f_i > 0$, and $f'' < 0$ in $(0, 1)$ for $i = 1, 2$. It follows from (3.2) that

$$h'''(\eta) = 2Q \int_{\xi}^{\eta} f_2''' h e^{-2Q\int_x^{\eta} f_1\, dt}\, dx + h'''(\xi) e^{2Q\int_{\xi}^{\eta} f_1\, dt}. \tag{3.5}$$

Let $h'''(0) = \mu$. We will prove that if $h(\eta)$ is nontrivial and satisfies (3.3) with $h(1) = 0$, then $h''(1)$ never becomes zero. This implies that the boundary value problem of linear ordinary differential equations (3.2) with (3.3)-(3.4) has only a trivial solution, as desired. Thus, there are several subcases to consider.

Case 1. $\lambda \geqq 0$ and $\mu > 0$. Initially, h, h', h'', and h''' are positive. Since $h(\eta)$ reaches its maximum somewhere in $(0, 1)$, there must be a point η' at which $h > 0$, $h' > 0$, and $h'' < 0$. Hence $h''' < 0$ at somewhere in between zero and η', say η''. However, once $h'''(\eta'') < 0$ then $h''' < 0$ as long as $h \geqq 0$ for $\eta > \eta''$, and so $h'' < 0$, $h' < 0$, and $h''' < 0$ when $h = 0$ at some $\eta = \eta_1 > \eta''$. In the proof of the following case, Case 3, we will see that if h has its next zero at $\eta = \eta_2$, then $h' > 0$, $h'' > 0$, and $h''' > 0$ at η_2, which will again lead to Case 1. Therefore, it is impossible that h'' does not equal zero whenever $h = 0$.

Case 2. $\lambda > 0$ and $\mu \leqq 0$. Initially, $h > 0$. Then, by (3.6), $h''' < 0$, and hence $h'' < 0$ as long as $h > 0$. If η'' is the next zero point of $h(\eta)$, then $h' < 0$, $h'' < 0$, and $h''' < 0$ at $\eta = \eta''$. This is the case that we have studied in Case 1, in which it has been proved that $h'' < 0$ at the next zero point of $h(\eta)$.

Case 3. $\lambda = 0$ and $\mu < 0$. Initially, $h'' < 0$, $h' < 0$, $h''' < 0$, and $h < 0$ for η close to zero. Suppose $h = 0$ somewhere. There would be a point, say η^* (< 1), at which $h'' > 0$, $h' > 0$, and $h < 0$. Also, there would be a $\xi^* \in (0, \eta^*)$ such that $h'''(\xi^*) > 0$. Then we see that $h''' > 0$ as long as $h \leqq 0$ for $\eta > \xi^*$ by (3.6). This means that $h' > 0$, $h'' > 0$, and $h''' > 0$ at $\eta = \eta_1$, the next zero of h. This is as in Case 1.

Case 4. $\lambda = 0$ and $\mu > 0$. This is the case where $h'' > 0$, $h' > 0$, and $h > 0$, initially, which is similar to Case 1.

In summary, we conclude that if $h(\eta)$ has only a finite number of zeros in $[0, 1]$, then $h'' = h = 0$ for $\eta > 0$ will never occur simultaneously. Suppose that $h(\eta)$ has infinitely many zeros in $[0, 1]$. Then, there must be a sequence $\{\eta_n\}$ $(n = 1, 2, \cdots)$ of the zero points of $h(\eta)$ such that $\eta_n \to \zeta \in [0, 1]$, and $\eta_n < \eta_{n+1}$, and such that $h' > 0$, $h'' > 0$, and $h''' > 0$ at η_n for $n = 1, 3, \cdots$, and $h' < 0$, $h'' < 0$, and $h''' < 0$ at η_n for $n = 2, 4, \cdots$. The continuity of h''' yields $h = h' = h'' = h''' = 0$ at ζ, therefore $h \equiv 0$, which is impossible. The conclusions of Theorem 2 now follow.

Theorem 3 implies that all the flows for the floating slot are two-celled. Our next theorem indicates that for some $Q > 0$ the number of solutions of (1)-(2) bifurcates as A varies. That is, in the case where $A = 2$ there is only one solution for any given $Q > 0$, i.e., only one β corresponds to one Q; and if $A = 1$ there are admitted multiple solutions for some $Q > 0$, i.e., there are more than one (at least three) β corresponding to one Q as depicted in [1]. Since, in fact, we have proved in the case where $A = 1$ that for any given $Q > 0$ there is admitted at least one β such that problem (1)-(2) has a convex solution [1], we only need to prove that there exist some $Q > 0$ for which there are admitted some β such that the corresponding solutions are nonconvex.

THEOREM 4. *Let $A = 1$ in (1). There exist some $Q > 0$ and corresponding β such that (1)-(2) has at least one nonconvex solution.*

Proof. We apply the same transformation used in § 1 again and obtain the following equation:

$$\varphi''' - [\varphi\varphi'' - \varphi'^2] = -\frac{\beta Q}{c^4} \tag{3.6}$$

with boundary conditions (1.3)-(1.4). Differentiating (3.6) once gives

$$\varphi'''' = \varphi\varphi''' - \varphi'\varphi''. \tag{3.7}$$

Consider the initial value problem (3.7) with initial conditions

$$\varphi(0) = \varphi''(0) = 0, \quad \varphi'(0) = \lambda, \quad \varphi'''(0) = \mu. \tag{3.8}$$

Our purpose is to find some pairs (λ, μ) such that problem (3.7)-(3.8) has nonconvex solutions. The proof of this theorem is contained in the following lemmas.

LEMMA 5. *$\varphi'''' < 0$ for all $\eta > 0$ as long as $\lambda\mu \neq 0$.*

Proof. Differentiating (2.7), we obtain that

$$\varphi^{(5)} = \varphi\varphi'''' - \varphi''^2; \tag{3.9}$$

hence

$$\varphi''''(\eta)^{-\int_0^\eta \varphi\, dt} = -\int_0^\eta \varphi''^2 e^{-\int_0^x \varphi\, dt}\, dx. \tag{3.10}$$

Since $\lambda\mu \neq 0$ implies that $\varphi'' \neq 0$ initially, $\varphi'''' < 0$ for all $\eta > 0$. This proves the lemma.

Remarks. It is seen from Lemma 5 that if $\lambda \leqq 0$ and $\mu < 0$, then $\varphi''' < 0$, $\varphi'' < 0$, $\varphi' < 0$, and $\varphi < 0$ for all $\eta > 0$. Also, if $\lambda > 0$ and $\mu < 0$, then φ''', $\varphi'' < 0$ for all $\eta > 0$; hence $\varphi = 0$ somewhere. This is a convex solution, which we found in [1].

LEMMA 6. *φ must reach zero with $\varphi'' < 0$ for some $\eta > 0$ in the following two cases:*

(I) *$\lambda > 0$ and $\mu > 0$.*

(II) *$\lambda < 0$ and $\mu > 0$.*

Proof. Case I. Initially, $\varphi, \varphi', \varphi'', \varphi''' > 0$, but $\varphi'''' < 0$. By (3.10), $\varphi^{(5)} < 0$ and hence φ'' must reach zero somewhere, say η_d, with $\varphi'''(\eta_d) < 0$ so long as $\varphi > 0$. Therefore, $\varphi''(\eta) < 0$ for $\eta > \eta_d$, which will lead $\varphi' = 0$ and φ'', $\varphi''' < 0$ at some $\eta_1 > \eta_d$. Thus φ must become zero at some $\eta = \eta_2$ with $\varphi''(\eta_2) < 0$.

Case II. Initially, $\varphi, \varphi' < 0$ but φ'', $\varphi''' > 0$. We claim that there must be a point $\eta = \xi$ where $\varphi''' = 0$. To see this, we suppose the contrary: that $\varphi''' > 0$ for all $\eta > 0$. Then φ' is concave up, so φ' becomes zero somewhere, say $\eta = \eta_3$. This yields $\varphi = 0$ with $\varphi'' > 0$ at some $\eta_4 > \eta_3$. Therefore, $\varphi^{(5)} \leqq -\varphi''^2 < 0$ for $\eta \geqq \eta_4$, and φ''' must be zero somewhere, which contradicts the above supposition. Let $\varphi'''(\eta_0) = 0$. We see that $\varphi'(\eta_0) > 0$ by $\varphi''(\eta_0) > 0$ and $\varphi''''(\eta_0) = -\varphi'(\eta_0)\varphi''(\eta_0) < 0$. Then $\varphi'''' < 0$ gives $\varphi''' < 0$ for all $\eta > \eta_0$, and φ reaches its first zero somewhere with $\varphi'' > 0$. Furthermore, φ must reach its second zero with $\varphi'' < 0$ due to $\varphi'''' < 0$ and $\varphi''' < 0$. This is what we want to prove. From the above proof we see that to get nonconvex solutions, $\varphi'''(0)$ must be greater than zero. Hence $\lambda^2 + \mu = -\beta Q/c^4 > 0$. This proves the following corollary.

COROLLARY. *Any nonconvex solution of the boundary value problem* (3)-(4) *corresponds to $\beta > 0$.*

In addition, we have found the three-cell flows for the floating slot analytically in the proof of Theorem 4, which occurs if $\lambda < 0$ and $\mu > 0$.

4. Existence of solutions for $\beta < 0$, and the limit of the convex solutions as $Q \to \infty$. As a complement to our previous paper [1], in this section we prove the existence of solutions for negative β for all $A \in [1, 2]$. Also, we study the limit of the convex solutions $g(x)$ of (3)-(4) as $Q \to \infty$.

THEOREM 5. *For given* $A \in [1, 2]$ *there exists a number* $Q_0 = Q_0(A)$ *such that if solutions of* (3)-(4) *are convex and* $Q > Q_0$, *then* $\beta < 0$. *Particularly,* $Q_0(2) \leqq 6(e^6 - 1)$.

Proof. Suppose that $\varphi(x)$ is a concave down solution. Then $\varphi'''' < 0$ by the corollary in § 2; hence $\varphi(1) < \lambda + \mu/6$ and $\mu > -6\lambda$. Thus $\beta = -(\lambda^2 + \mu)/Q$. To prove the theorem we only need to show that $\lambda > 6$ for large Q. Assume the contrary, that is, that $0 < \lambda \leqq 6$, and then $-36 < \mu < 0$ for $Q_n \uparrow +\infty$. There would be subsequences of $\{\lambda_n\}$ and $\{\mu_n\}$, say $\{\lambda'_n\}$ and $\{\mu'_n\}$, such that $\lim \lambda'_n = \bar{\lambda}$ and $\lim \mu'_n = \bar{\mu}$ as $n \to \infty$. Since $\bar{\lambda}$ and $\bar{\mu}$ are bounded, the corresponding solution $\varphi(x; \bar{\lambda}, \bar{\mu})$ would have the finite second derivative at $x = 1$, but $\varphi''(1; \lambda'_n, \mu'_n) = -Q'_n \uparrow -\infty$. This contradicts the theory that the solutions continuously depend on the initial conditions. Therefore, there exists a $Q_0 = Q_0(A)$ such that if $Q > Q_0$, then $\beta < 0$.

In particular, let $A = 2$ and $g(x) = \varphi(x)Q$ in (3). Then the equation becomes

$$\varphi'''' = 2\varphi\varphi''', \tag{4.1}$$

with boundary conditions

$$\varphi(0) = \varphi''(0) = \varphi(1) = 0, \qquad \varphi''(1) = -Q. \tag{4.2}$$

Again let $\varphi'(0) = \lambda$ and $\varphi'''(0) = \mu$. Rewriting (4.1) as

$$\varphi'''(x) = \mu e^{2\int_0^x \varphi\, dt} \tag{4.3}$$

it is observed by $\mu < 0$ and $0 < \varphi < \lambda x$ for $x \in (0, 1)$ that

$$-Q = \mu \int_0^1 e^{2\int_0^x \varphi\, dt}\, dx > \mu \int_0^1 e^{\lambda x}\, dx = \frac{\mu(e^\lambda - 1)}{\lambda}. \tag{4.4}$$

Then

$$\mu < -\frac{\lambda Q}{e^\lambda - 1}. \tag{4.5}$$

On the other hand, since $\varphi'''' < 0$ for $x \in (0, 1)$, $\varphi(1) < \lambda + \mu/3!$. This gives $\mu > -6\lambda$. Combining this inequality with (4.5), we obtain $6(e^\lambda - 1) > Q$, namely, $\lambda > \ln(1 + Q/6)$.

Note that β in equation (1) is equal to $-(\lambda^2 + \mu)/Q$. Then $\lambda^2 + \mu > \lambda^2 - 6\lambda$. Since the hypothesis on Q gives $\ln(1 + Q/6) > 6$, we conclude that $\lambda^2 + \mu > 0$, i.e., $\beta < 0$.

Combining the corollary in § 3, we complete the proof.

Remark. It is seen from the proof of Theorem 5 that $\lambda \to \infty$ as $Q \to \infty$, i.e., $\lambda = O(1)$ for large Q. For $A = 2$ we have proved $\lambda > \ln Q$ as $Q \to \infty$. The next theorem will show $\lambda/Q \to 0$ as $Q \to \infty$ for the convex solutions.

THEOREM 6. *Suppose that* $1 \leqq A \leqq 2$ *and that* $g(x)$ *are concave down, and solve* (3) *with the boundary conditions* (4). *As* $Q \to \infty$, *then* $g(x) \to 0$ *and* $g'(x) \to 0$ *uniformly in* $[0, 1]$, *and* $g''(x) \to 0$ *uniformly in* $[0, x_0]$ *for any* $x_0 \in (0, 1)$.

Proof. Recall that the concave down solutions we have found satisfy $g'''' < 0$ in $(0, 1)$, which has been proved in § 2. We will prove that for any given real sequence $\{Q_n\}$ with $Q_n \to \infty$ there exists at least a subsequence $\{Q'_n\}$ such that $g(x; Q'_n) \to 0$ and $g'(x; Q'_n) \to 0$ uniformly in $[0, 1]$, and $g''(x; Q'_n) \to 0$ uniformly in $[0, x_0]$ for any $x_0 < 1$. Let $g'(0) = \sigma$ and $g'''(0) = \kappa$. Rewrite (3) as

$$\frac{g'''}{Q} = Agg'' - g'^2 + \sigma^2 + \frac{\kappa}{Q}. \tag{4.6}$$

Integrating (4.6), we obtain

$$\frac{g''}{Q} = Agg' - (A+1)\int_0^x g'\, dt + \left(\sigma^2 + \frac{\kappa}{Q}\right)x. \tag{4.7}$$

The fact $g'''' < 0$ and the boundary conditions imply that $-1 \leqq g'' \leqq 0$, $|g'| \leqq 1$, and $0 \leqq g \leqq 1$ for all $x \in [0, 1]$. Thus the function sequence $\{g'(x; Q_n')\}$ is equicontinuous. By the Arzelà-Ascoli theorem there exists a subsequence $\{g_{n_k}'\}$ that converges uniformly to a continuous function $h(x)$ on $[0, 1]$. Similarly, $\{g_{n_k}\}$ contains a subsequence, say $\{g_{n_k}\}$ for simplicity, that converges uniformly to a function $k(x)$ on $[0, 1]$. Thus $k'(x) = h(x)$. In brief, we drop the indices n_k in the rest of the proof. It is known that $g(0) = \sigma$ and $g'''(0) = \kappa$ are bounded because of negative g'' and g'''' and $g''(1) = 1$. Hence $\lim \sigma = \sigma_0$ and $\lim k/Q = 0$ (appropriate subsequence of n_k) as $Q \to \infty$. Set $Q \to \infty$ in (4.7). This yields

$$Akk' - (A+1)\int_0^x k'^2\, dt + \sigma_0 x = 0 \qquad (0 \leqq x \leqq 1). \tag{4.8}$$

Suppose that there is a point $x_0 \in (0, 1)$ at which $k \neq 0$. Then $k \neq 0$ in an open subinterval I of $(0, 1)$. By (4.8) we see that on I

$$k' = \left[(A+1)\int_0^x k'^2\, dt - \sigma_0 x \right] \cdot (Ak)^{-1}. \tag{4.9}$$

Then on the interval I, k'' and k''' exist and satisfy (differentiating (4.8) twice)

$$Ak''' = 0. \tag{4.10}$$

So $k''' = 0$ and $k = ax^2 + bx + c$ on I, where a, b, and c are some constants. Substituting k into (4.8), we find $a = 0$. Therefore, k is a linear function on I. Since k is also convex, $k(x_0) \neq 0$ implies that $I = (0, 1)$. Hence the homogeneous boundary conditions of g lead to $b = c = 0$, and $k \equiv k' \equiv 0$.

We have proved that g and g' uniformly converge to zero on $[0, 1]$. Next we will show that $\{g''\} \to 0$ uniformly on $[0, 1-\delta]$ for any small $\delta > 0$. Since g'' are convex, as $Q \to \infty$,

$$0 > g'' > g''(1-\delta) > \frac{g'(1-\delta/2) - g'(1-\delta)}{\delta/2} \to 0.$$

The proof of Theorem 6 is now complete.

Remark. Note that $\lambda = Q\sigma$ and $\mu = Q\kappa$ in the proofs of Theorems 5 and 6. We see that $\lambda \to \infty$ and $\lambda/Q \to 0$ as $Q \to \infty$. In the case $A = 2$, we have proved $\ln Q < \lambda < Q$. Also, the conclusion of Theorem 6 agrees with the numerical results [4].

REFERENCES

[1] C. Lu, N. D. Kazarinoff, J. B. McLeod, and W. C. Troy, *Existence of solutions of the similarity equations for floating rectangular cavities and disks*, SIAM J. Math. Anal., 19 (1988), pp. 1119-1126.

[2] S. P. Hastings, *An existence theorem for a problem in boundary layer theory*, Arch. Rational Mech. Anal., 33 (1969), pp. 103-109.

[3] S. P. Hastings, C. Lu, and Y-H. Wan, *A three-parameter shooting method as applied to a problem in combustion theory*, Phys. D, 19 (1986), pp. 301-306.

[4] N. D. Kazarinoff, *Numerical results on the similarity equations*, private communication, 1984.

[5] W. N. Gill, N. D. Kazarinoff, and J. D. Verhoeven, *Convective diffusion in zone refining of low Prandtl number liquid metals and semiconductors*, in Integrated Circuits: Chemical and Physical Processing, P. Stroeve, ed., Amer. Chem. Soc. Symposium Series, 290, 1985, pp. 47-69.

[6] W. N. Gill, N. D. Kazarinoff, C. C. Hsu, M. A. Noack, and J. D. Verhoeven, *Thermocapillary driven convection in supported and floating driven convection*, Adv. Space Res., 4 (1984), pp. 15-22.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 729-756, May 1990
© 1990 Society for Industrial and Applied Mathematics
010

FLAT CONNECTIONS AND SCATTERING THEORY ON THE LINE*

D. H. SATTINGER†

Abstract. The scattering theory of $n \times n$ first-order systems on the line is formulated in terms of a flat connection on a vector bundle over $R \times P_1(C)$. The relation of the scattering data to a set of transition matrices is discussed. The scattering transform is obtained as a sectionally holomorphic gauge transformation. The winding number constraints of Bar-Yaacov ["Analytic properties of scattering and inverse scattering for first order systems," Ph.D. thesis, Yale University, New Haven, CT, 1985] and Beals and Coifman [*Comm. Pure Appl. Math.*, 37 (1984), pp. 39-90] on the scattering data are shown to be a necessary condition for the diagram for the transition matrices to commute. The transition matrices are reconstructed from scattering data with multiple poles by solving a sequence of triangular factorization and Riemann-Hilbert problems. The inverse scattering problem is formulated as a system of singular integral equations and reduced to a Fredholm system by Plemelj's method for a special class. A formulation of the dressing method for $n \times n$ hierarchies in terms of sectionally holomorphic gauge transformations is given.

Key words. scattering theory, flat connections

AMS(MOS) subject classification. 34A55

1. Introduction.

Consider the matrix linear first-order differential operator

$$D_x(z, Q) = \frac{d}{dx} - zJ - Q \tag{1.1}$$

where J is a diagonal matrix and Q is an off-diagonal matrix (viz. $Q_{jj} = 0$). More generally, we may consider the case where J and Q are elements of a semisimple Lie algebra $\mathbf{g}$, with J in the Cartan subalgebra. Such differential operators arise as isospectral operators in a large class of completely integrable systems of partial differential equations. The solution of the forward and inverse scattering problem for (1.1) is now fairly complete, due to the work of Beals, Coifman, Deift, Tomei, and Zhou [2], [5], [16].

Beals and Coifman solved the inverse scattering problem for potentials Q for which the scattering data consists of simple poles with residues of a specified type (to be discussed below). Such potentials are generic in the sense that they are dense in the topology of $L_1(R)$; but the class is not invariant under Bäcklund transformations [12], [17].

The forward and inverse scattering problems can be formulated naturally in the geometric language of vector bundles. Using this approach, we extend the analysis of the scattering problem given in [2], [5] to a more general class of scattering data. We use the construction of scattering data given by Zurkowski [17], which is invariant under Bäcklund transformations. We show how the scattering data in the form given by Zurkowski is related to the transition matrices of the bundle, and show how to reconstruct the transition matrices from the scattering data. This entails the solution of a sequence of scalar Riemann-Hilbert problems and involves the so-called winding number constraints introduced by Bar-Yaacov [1]. We derive the winding number constraints as a consequence of the requirement that the transition matrices of a certain principal bundle must intertwine and give them in a simplified, explicit form.

* Received by the editors June 22, 1988; accepted for publication (in revised form) June 26, 1989. This research was partially supported by National Science Foundation grant DMS-87-02578.

† School of Mathematics, University of Minnesota, 127 Vincent Hall, Minneapolis, Minnesota 55455.

In [2] the inverse scattering problem is formulated as a Riemann-Hilbert problem for a wave function $m(x, z)$ with poles in the complex plane. First, a rational approximation to the scattering data is constructed, and then a small norm problem for the remainder is solved. This leads to a wave function m with, in general, different poles than those required of the solution. An algebraic problem must then be solved to obtain the correct poles. In [5], the inverse scattering problem is reduced to inverting a linear operator with the structure 1+"small"+"compact," for which the Fredholm alternative applies. For certain nth-order self-adjoint scalar equations, the existence and uniqueness of the solution is accomplished by the use of the so-called "vanishing lemma" developed by Deift, Tomei, and Trubowitz [6], [7].

We present an alternative approach to the inverse scattering problem in § 6. There is a large classical literature on Riemann-Hilbert problems, going back to Plemelj. We show how a variant of Plemelj's method can be used to reduce the problem to a Fredholm integral equation in the case $J^* = -J$. The Riemann-Hilbert problem of inverse scattering involves the space variable x as a parameter, and the solution must satisfy certain "radiation conditions" as $x \to \pm\infty$. This condition is easily verified for the Fredholm integral equation that we obtain.

In general, we cannot guarantee that the Fredholm problem is solvable; however, in certain cases a general result of this nature can be established. For the nth-order self-adjoint scalar case, the vanishing lemma is used to prove that the kernel is trivial. For first-order $n \times n$ systems, the scattering data arising from skew-Hermitian potentials ($J^* = -J$ and $Q^* = -Q$) has certain symmetry properties that guarantee the unique solvability of the inverse scattering problem [13].

2. Scattering data for $n \times n$ systems. In this section we review the basic results of scattering theory for one-dimensional systems. We assume throughout that the eigenvalues of J are distinct with nondecreasing real parts: $\operatorname{Re} \lambda_1 \leqq \operatorname{Re} \lambda_2 \leqq \cdots \leqq \operatorname{Re} \lambda_n$. We look for a solution to the eigenfunction problem $D_x(z, Q)\Psi(x, z) = 0$ in the form $\Psi(x, z) = m(x, z)\, e^{xzJ}$. Then m satisfies the differential equation

$$\frac{\partial m(x, z)}{\partial x} = z[J, m] + Q(x)m. \tag{2.1}$$

Equation (2.1) is supplemented by the "boundary" condition

$$\sup_x |m(x, z)| < +\infty \tag{2.2}$$

and by the asymptotic condition that $m \to 1$ as $z \to \infty$.

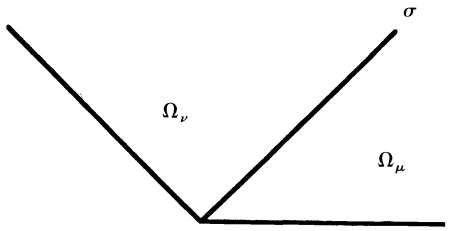


FIG. 2.1. *Sectors of analyticity in the complex plane.* $\sigma = \mu^{-1}\nu$.

To carry out the analysis of (2.1), (2.2), we introduce the following sectors in the complex plane. Let z be a complex number for which $\operatorname{Re} z(\lambda_j - \lambda_k) \neq 0$ and let ν be a permutation of $\{1, 2, \cdots, n\}$. Define

$$\Omega_\nu = \{z \mid \operatorname{Re} z(\lambda_{\nu(j)} - \lambda_{\nu(k)}) < 0 \text{ for } j < k\}.$$

Of course, not every permutation will correspond to a sector. Let Σ_σ, $\sigma = \mu^{-1}\nu$, be the ray running from zero to infinity that separates two sectors, with Ω_ν on the left, and Ω_μ on the right, as shown in Fig. 2.1. The right action of the permutation group on $n \times n$ matrices is given by $JT_\nu = \nu^{-1} J \nu$, where T_ν is the $n \times n$ permutation matrix associated with ν. Then $T_\mu T_\nu = T_{\mu\nu}$. If $J = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$, then $J_\nu = JT_\nu = \operatorname{diag}(\lambda_{\nu(1)}, \cdots, \lambda_{\nu(n)})$.

We call the *standard* representation that in which the real parts of the eigenvalues of J are nondecreasing. By the *ν representation* of a matrix U we mean the matrix $UT_\nu = \nu^{-1} U \nu$. Below, we will use the phrase *ν-lower* to indicate that a matrix is lower triangular in the ν representation, etc.

THEOREM 2.1 [2]. *Let Q belong to $L_1(R)$; then there exists a unique solution $m(x, z)$ of* (2.1), (2.2) *such that* (i) *$m(x, z)$ is meromorphic in each of the sectors Ω_ν;* (ii) *$m(x, z) \to 1$ as $z \to \infty$ in Ω_ν or as $x \to -\infty$ for fixed $z \in \Omega_\nu$;* (iii) *$m(x, z)$ is bounded on $-\infty < x < \infty$ for regular values of z. In each sector Ω_ν, m has continuous boundary values on $\partial\Omega_\nu$, except at possible limit points of the poles on the boundary.*

The poles of m correspond to the bound states in the 2×2 case. For general potentials in L_1, the poles of m constitute a bounded, discrete set Z; but for a dense open subset of potentials in L_1, Z is finite and the poles of m are simple [2]. In this paper we treat general poles of m, with none on the rays Σ. The case of an infinite number of poles (clustering at a point in Σ) has been treated by Zhou [16]. In this paper we will indicate another method of treating an infinite number of poles.

In [2] and [5], the scattering data of Q is given in terms of the singularities of m. These consist of the location of its poles, the singular parts of m at these poles, and the jump conditions of m across the rays Σ_σ. The jump conditions comprise the continuous component of the scattering data and are obtained as follows. Let $m_\nu(x, \xi)(m_\mu(x, \xi))$ be the limit of m as $z \to \xi \in \Sigma_\sigma$ from the regions $\Omega_\nu(\Omega_\mu)$. It is easily verified by direct substitution that $w = m_\mu^{-1} m_\nu$ satisfies the differential equation

$$D_\xi w \equiv \left(\frac{d}{\partial z} - \xi \operatorname{ad} J\right) w = 0, \qquad \xi \in \Sigma_\sigma.$$

The solutions of this equation are of the form $\exp^{x\xi J} V_\sigma(\xi) \exp^{-x\xi J}$. We denote such an expression by $V_\sigma(\xi)^x$, so m satisfies the jump conditions $m_\nu(x, \xi) = m_\mu(x, \xi) V_\sigma(\xi)^x$, $\xi \in \Sigma_\sigma$. From now on we will denote the collection of matrices $\{V_\sigma\}$ simply by V.

At an isolated pole, say z_j, of m, the wave function can be factored (cf. [12], [17]) as a product $m = \eta_j(x, z)(1 + L_j(z))^x$, where η_j satisfies (1.2) in Ω_ν, tends to 1 as $x \to -\infty$ or as $z \to \infty$ for $x < 0$, and L_j is strictly ν-lower and rational in z, with an isolated pole at z_j. If m has a finite number of isolated poles in Ω_ν, we can repeat the process a finite number of times until m has been factored in the form $m = \eta(x, z)(1 + L(z))^x$, where $L(z)$ is strictly ν-lower, rational in z, and tending to zero as z tends to infinity, and η is holomorphic in Ω_ν and satisfies the usual asymptotic conditions at $-\infty$ for $x < 0$. Such a factorization is unique. If m has two such factorizations, $m = \eta(1 + L)^x = \eta'(1 + L')^x$, then $\eta'^{-1}\eta = [(1 + L')(1 + L)^{-1}]^x$; but the left side is analytic in Ω_ν and the

right side is meromorphic in the entire complex plane with its only possible singularities in Ω_ν. Therefore the right side is in fact entire, and since it tends to 1 as $z \to \infty$ (for $x < 0$), it is identically 1. Therefore $\eta = \eta'$ and $L = L'$.

We may repeat the same arguments for $x > 0$. Since, however, m tends to a diagonal matrix δ_ν as $x \to \infty$ [2], we obtain the unique factorization $m = \rho_\nu(1 + U_\nu)^x \delta_\nu$ in the sector Ω_ν for $x > 0$, where ρ_ν is an analytic solution of (2.1) and U_ν is a strictly ν-upper triangular matrix that is rational in z and tends to zero as $z \to \infty$. Summarizing, we have the following theorem [12], [17].

THEOREM 2.2. *The wave function $m(x, z)$ can be uniquely factored in $\Omega_\nu \times \{x \leqq 0\}$ as $m = \eta_\nu(1 + L_\nu)^x$ where*

(i) *η_ν is an analytic solution of* (2.1) *in Ω_ν that tends to* 1 *as $x \to -\infty$ or as z tends to infinity for $(z, x) \in \Omega_\nu \times \{x \leqq 0\}$, and*

(ii) *L_ν is strictly ν-lower, rational in z in Ω_ν, analytic in Ω_ν^c, and tends to zero as $z \to \infty$.*

Similarly, in $\Omega_\nu \times \{x \geqq 0\}$, m can be uniquely factored as $m = \rho_\nu(1 + U_\nu)^x \delta_\nu$, where

(iii) *ρ_ν is an analytic solution of* (2.1) *in Ω_ν that tends to* 1 *as $x \to \infty$ or as $z \to \infty$ for $(z, x) \in \Omega_\nu \times \{x \geqq 0\}$, and*

(iv) *U_ν is strictly ν-upper, rational in z in Ω_ν, analytic in Ω_ν^c, and tends to zero as $z \to \infty$.*

(v) *δ_ν is a diagonal matrix, meromorphic in Ω_ν, that tends to* 1 *as $z \to \infty$ in Ω_ν.*

From the two representations $m = \eta_\nu(1 + L_\nu)^x = \rho_\nu(1 + U_\nu)^x \delta_\nu$, we obtain the matrix $S_\nu(z)$ from $S_\nu(z)^x = \rho_\nu^{-1}\eta_\nu = (1 + U_\nu)^x \delta_\nu[(1 + L_\nu)^{-1}]^x$. We are now in a position to define the scattering data (cf. [17]).

DEFINITION 2.3. The scattering data of Q is s.d. $(Q) = \{V; z_j; L_\nu\}$, where $\{z_j\}$ are the totality of poles of m in $\Omega \backslash \Sigma$, L_ν is rational in z in Ω_ν, analytic for $z \in \Omega_\nu^c$ and tends to zero as $z \to \infty$.

The matrix $V(\xi)$ is defined on the set of rays Σ. The matrices V and L_ν are not completely arbitrary, but must satisfy a number of constraints that will be discussed in the course of this article. For one thing, the $k \times k$ lower minors, $\det_k^-(V)$, must be identically equal to 1; and a set of "winding number constraints" must be satisfied by the upper minors $\det_k^+(V)$ and the singular parts L_ν. The precise relationship is given in Theorem 4.8.

The relationship between the V and Q is much the same as the relation between a function and its Fourier transform. (As is well known by now, the scattering transform is a nonlinear analogue of the Fourier transform.) If the derivatives $D^j Q \in L_1(R)$ for $1 \leqq j \leqq k$, then $\xi^k(V(\xi) - 1) = o(1)$ as $\xi \to \infty$; if $x^k Q \in L_1$, then $V \in C^k(\Sigma)$ and $D^j(V - 1) \to 0$ as $\xi \to \infty$ for $0 \leqq j \leqq k$ (cf. [2, Thm. E]).

A particularly convenient space of potentials to work with is the Schwartz space $S(R)$ of C^∞ rapidly decreasing functions. If $Q \in S(R)$, then each $V \in S(\Sigma)$ (cf. [2] for the behavior of V in the neighborhood of the origin). In this paper we restrict ourselves to potentials and scattering data in the Schwartz class, since regularity questions are not the issue here. For a treatment of more general data, see [16].

So far we have shown how to obtain the scattering data beginning with the wave function m constructed by Beals and Coifman. In their picture, the singularities in m constitute the scattering data. In the bundle view, however, the singularities of m arise from gluing together different coordinate patches of a principal bundle. We now present an alternative construction of m that is consistent with the bundle viewpoint to be developed in § 3.

THEOREM 2.4. [17] *There exist solutions $\eta_\nu(x, z)$ and $\rho_\nu(x, z)$ of* (2.1) *in $\Omega_\nu \times (-\infty, a]$ and $\Omega_\nu \times (b, \infty]$, respectively, that are analytic for $z \in \Omega_\nu$, continuous in the*

closure, and satisfy the asymptotic conditions $\eta_\nu \to 1$ *as* $x \to -\infty$, $\rho_\nu \to 1$ *as* $x \to \infty$ *for* $z \in \Omega_\nu$; *and* $\eta_\nu \to 1$ *as* $z \to \infty$ *for* $x \leqq a$ *and* $\rho_\nu \to 1$ *as* $z \to \infty$ *for* $x \geqq b$.

For any complex z, let Π_ε^z denote the projections onto the positive ($\varepsilon = 1$), negative ($\varepsilon = -1$), and null ($\varepsilon = 0$) subspaces of the operator Re z ad J; that is,

$$(\Pi_\varepsilon^z a)_{jk} = a_{jk} \quad \text{if sgn Re } z(\lambda_j - \lambda_k) = \varepsilon, \qquad (\Pi_\varepsilon^z a)_{jk} = 0 \quad \text{otherwise}$$

(where we define sgn $0 = 0$). We obtain a wave function η_ν that is normalized at $-\infty$ by solving the integral equation

$$\eta_\nu = 1 + \int_{-\infty}^{x} (\Pi_-^z + \Pi_0^z)\, e^{(x-y)zJ} Q \eta_\nu\, e^{-(x-y)zJ}\, dy - \int_x^a \Pi_+^z\, e^{(x-y)zJ} Q \eta_\nu\, e^{-(x-y)zJ}\, dy. \tag{2.3}$$

For $z \in \Omega_\nu$, (2.3) is a Fredholm integral equation, and the solutions will have poles in z. For sufficiently negative a, however, the norm of the integral operator will be less than 1 on $(-\infty, a]$, and (2.3) can be solved by successive approximations. This gives a solution of (2.1) on $\Omega_\nu \times (-\infty, a]$ that tends to 1 as $x \to -\infty$. The solution for $x \geqq a$ is obtained by solving the differential equation on $x \geqq a$ using the data $\eta_\nu(a, z)$ obtained from the solution of the integral equation on $(-\infty, a)$. By this method we obtain a solution that tends to 1 as $x \to -\infty$, and is analytic for $z \in \Omega_\nu$. For $x \leqq a$, the solution will tend to 1 as $z \to \infty$ in Ω_ν, *but this asymptotic behavior does not hold for* $x > a$.

Similarly, ρ_ν is obtained by solving a Fredholm integral equation normalized at $+\infty$. This solution is analytic in Ω_ν and tends to 1 as $z \to \infty$ for $x \geqq b$. In general, $a < b$, the two solutions will not cover the entire real line, and additional solutions on intermediary intervals will be required. (Such a situation was, in fact, discussed in the original method of [2].) If, however, the norms of the integral operators in (2.3) are less than 1 over the entire interval $(-\infty, 0]$ and $[0, \infty)$, then we may cover the entire line with two patches for each Ω_ν. Of course, in the case where m is known to have only a finite number of poles, η_ν and ρ_ν may be constructed on $(-\infty, 0]$ and $[0, \infty)$, respectively, from m itself, using the procedure outlined prior to Theorem 2.2. For now, let us stick to the case where only two patches are needed.

The solutions η and ρ are not uniquely determined. Given any two solutions on $x < 0$, η_1 and η_2, say, it is easily seen that $D_z \eta_1^{-1} \eta_2 = 0$, so that $\eta_1 = \eta_2 W(z)^x$ for some $W(z)$ analytic in Ω_ν. From the asymptotic conditions $\eta_j \to 1$ as $z \to \infty$ we conclude that $W(z) = 1 + L(z)$, where L is strictly ν-lower triangular for $z \in \Omega_\nu$.

Since ρ_ν and η_ν are both solutions of (2.1), $\eta_\nu = \rho_\nu S_\nu(z)^x$, where S_ν is analytic in Ω_ν. In fact, by the argument above, $D_z \rho_\nu^{-1} \eta_\nu = 0$, hence $\rho_\nu^{-1} \eta_\nu = S_\nu(z)^x$ for some matrix $S_\nu(z)$.

The wave function m in each sector Ω_ν can be constructed from ρ_ν, η_ν, and S_ν. We perform a triangular factorization of S_ν, namely,

$$S_\nu = (1 + U_\nu)\delta_\nu (1 + L_\nu)^{-1}, \tag{2.4}$$

where $U_\nu (L_\nu)$ is ν-upper (lower) triangular, and δ_ν is diagonal. Writing (2.4) in the form $S_\nu(1 + L_\nu) = (1 + U_\nu)\delta_\nu$, it can be seen that the factorization requires solving a system of linear equations for each of the $n(n-1)/2$ zero entries of the matrix $(1 + U_\nu)$. The factorization can be carried out provided that the lower principal minors of S_ν do not vanish. That is, let $\Delta_j^- = \det_j^- S_\nu$, where

$$\det_j^- S = \det \begin{bmatrix} S_{n-j+1,n-j+1} & \cdots & S_{n-j+1,n} \\ S_{n,n-j+1} & \cdots & S_{n,n} \end{bmatrix}.$$

Then the factorization (2.4) can be carried out provided none of the $\Delta_j^-(z)$, which are holomorphic in Ω_ν, vanish.

The zeros of these minors give precisely the eigenvalues in the 2×2 case. In the general case, the minors tend to 1 as $z\to\infty$ in Ω_ν. We take $m_\nu=\eta_\nu(1+L_\nu)^x=\rho_\nu(1+U_\nu)^x\delta_\nu$ for all values of $z\in\Omega_\nu$ for which the lower minors of S_ν do not vanish. We claim m is bounded in x for these regular values of z. The entries of $(1+L_\nu)^x$ are $(\delta_{jk}+L_{jk})\exp\{(\lambda_j-\lambda_k)xz\}$. For $z\in\Omega_\nu$, working in the ν representation,

$$\exp\{(\lambda_j-\lambda_k)xz\}\to 0 \quad \text{as } x\to-\infty \quad \text{for } j>k;$$

but $L_{jk}=0$ for $j<k$, so $(1+L_\nu)^x\to 1$ as $x\to-\infty$. Similarly, $(1+U_\nu)^x\to 1$ as $x\to\infty$. Therefore m is bounded on the entire real axis for all z for which the factorization (2.4) may be performed. The poles of m arise precisely at those z_j which the factorization breaks down.

3. The vector bundle viewpoint. Let us now turn to the formulation of the above picture in terms of sections of a *vector bundle.* We define a base space $B=R\times P_1(C)$ consisting of the entire real line crossed with the Riemann sphere and construct a trivial vector bundle over B by attaching to each point (x, z) in B the n-dimensional complex vector space C^n. The sections of this bundle are functions $v(x, z)$ taking values in C^n. We denote the bundle by E.

We have constructed a global basis, or frame, for E, namely, the constant sections $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n$. It will be useful, however, to consider other choices of frames. For example, if $g(x, z)$ is a matrix-valued function on B we obtain a new frame $\mathbf{e}_j(x, z)=g(x, z)\mathbf{e}_j$. Such a transformation g is called a *gauge* transformation. It is simply a change of basis in the fibers which varies from point to point over the base space. Associated with each frame is a matrix. The matrix associated with the constant frame $\mathbf{e}_1, \cdots, \mathbf{e}_n$ is the identity matrix. We will call this the standard frame. Clearly, $g(x, z)$ is the matrix associated with the frame $\mathbf{e}_j(x, z)=g(x, z)\mathbf{e}_j$, so the columns of g constitute the basis sections of the new frame. We may also refer to a frame as a gauge.

Now a differential operator $D_x=d/dx-U$, where U is an $n\times n$ matrix, may be interpreted as describing how sections of E vary with x. It is natural to ask how D_x transforms under a gauge transformation g. Let ψ and ψ' denote the coordinates of a section in two different frames related by a gauge transformation g; thus, we write $\psi'=g\psi$. Similarly, let D_x and $D'_x=d/dx-U'$ denote the differential operator in the two different frames (or gauges). Then we must have $(D_x\psi)'=D'_x\psi'$, or $gD_x=D'_xg$. It is not hard to see that this transformation law leads to the differential equation $dg/dx+gU-U'g=0$; in particular, the gauge transformation that gauges $d/dx-zJ-Q$ to $d/dx-zJ$ satisfies (2.1).

So far, we have said nothing about the dependence of the frame on z, or its behavior as x or $z\to\infty$. Since $Q\to 0$ as $|x|\to\infty$, it is natural to ask that the frame tend to the standard frame as $|x|\to\infty$; that is, that it in some sense be a perturbation of the standard frame. To discuss the z dependence, we introduce the Cauchy-Riemann operator $D_{z^*}=\partial^*=d/dz^*$. The operator D_{z^*} tells us how sections vary with z (in the "anti-holomorphic" direction). The pair of operators D_x and D_{z^*} form the components of a *connection.* Together, they tell us how sections vary along curves in the base space. Let us define $D\psi=(\partial\psi/\partial x-zJ\psi-Q\psi)\,dx-D_{z^*}\psi\,dz^*$. The equation $D\psi=0$ defines parallel translation of the section ψ along a curve in the base space.

Note that $[D_x, D_{z^*}]=0$, since D_x depends holomorphically on z. This means that, at least locally, it is possible to construct a frame g in which the connection takes the simpler form $\{d/dx-zJ, D_{z^*}\}$. All we need do is construct holomorphic solutions of (2.1) in some open neighborhood in B. Although E is trivial and the connection is flat, it is, however, not possible to construct a global frame that trivializes the connection.

Roughly speaking, the obstructions to such a trivialization constitute the "scattering data" for the problem.

To explain this remark, let us see how D_{z^*} transforms under g. We have

$$D_{z^*}g = g\frac{\partial}{\partial z^*} + \left(\frac{\partial}{\partial z^*}g\right) = g\left(\frac{\partial}{\partial z^*} + \mu\right),$$

where $\mu = g^{-1}\,\partial g/\partial z^*$. Note that $\partial g/\partial z^* = 0$ wherever g is analytic. For gauge transformations g with isolated singularities (for example, the wave function constructed in Theorem 2.1), $\partial g/\partial z^*$ is defined as the distribution

$$\partial^* g(\psi) = -\int\int g(\partial^*\psi)\, dz \wedge dz^*,$$

where ψ is a smooth (C^∞) section. With this definition it is not hard to see that if g is sectionally analytic in the neighborhood of a smooth oriented curve C, then $\partial^* g = g_+(\xi) - g_-(\xi)$, where $\pm$ refer to the limits of g on C consistent with its orientation. We return to the equation $\mu = g^{-1}\,\partial^* g$ below, and show that it, in fact, is a Riemann–Hilbert problem for the inverse problem.

For now, let us return to our discussion of the gauge transformation and the construction of a new frame. We wish to show how to construct a frame that tends to the standard frame as $|x|$ and z tend to infinity. Associated with the bundle E is the (trivial) principal bundle $B \times G$, where G is the Lie group with Lie algebra $\mathbf{g}$. The gauge transformations are sections of this principal bundle. As we explained above, the columns of a gauge transformation g constitute a frame (or gauge) for the vector bundle E.

Let $S_0(B, G)$ be the collection of all frames that are normalized to tend to the identity as $|x|$ or z tend to infinity. We construct sections of $S_0(B, G)$ that are solutions of the (2.1), (2.2) as follows. Choose a covering of the base space by the sets $\Omega_\nu^\pm = \Omega_\nu \times \{\pm x \geqq 0\}$, and *assume* (compare the comments below) we can construct a collection of local patches η_ν and ρ_ν that are defined and analytic (in z) and tend to 1 as $|x|$ and z tend to infinity in $\Omega_\nu^\pm$, respectively. The gauge transformations η_ν and ρ_ν constitute local frames in $\Omega_\nu^\pm$. At $\partial\Omega_\nu^\pm$ the frames must be matched by transition matrices. The transition matrices between η_ν and ρ_ν on $\Omega_\nu \times \{x = 0\}$ are S_ν, and the transition matrices for η_ν and ρ_ν are denoted by N_σ and R_σ, respectively. A global frame is now obtained in $C\backslash\Sigma$ by connecting these patches by the relevant transition matrices:

$$\begin{aligned} &\eta_\nu = \eta_\mu N_\sigma^x \quad \text{across } \Sigma_\sigma \times \{x \leqq 0\}, \\ &\rho_\nu = \rho_\mu R_\sigma^x \quad \text{across } \Sigma_\sigma \times \{x \geqq 0\}, \\ &\eta_\nu = \rho_\nu S_\nu^x \quad \text{on } \Omega_\nu \times \{x = 0\}. \end{aligned} \tag{3.1}$$

Comparing the jump conditions for g with those for η and ρ, we find that

$$N_\sigma = (1 + L_\mu)\, V_\sigma (1 + L_\nu)^{-1} \quad \text{and} \quad R_\sigma = (1 + U_\nu)\delta_\nu V_\sigma \delta_\mu^{-1}(1 + U_\mu)^{-1}.$$

An immediate necessary consequence of the relations (3.1) is the identity

$$R_\sigma S_\nu = S_\mu N_\sigma. \tag{3.2}$$

We show in § 4 that (3.2) leads directly to the winding number constraints on the scattering data [1], [2]. The identity (3.2) is a statement that the diagram of the transition matrices commutes.

We must still construct the frame on the set Σ. The frame for the vector bundle is singular in that the rays of Σ must be considered separately in the covering of the base space. This step is a little unorthodox from the point of view of vector bundles, since the set Σ is a one-dimensional subvariety of the base space B; but it is dictated by the analytic properties of the solutions of the differential equation.

When $J^* = -J$, the set Σ reduces to the real line, and solutions of (2.1) for real z (denoted by ξ), which tend to 1 as x tends, respectively, to $\pm\infty$ can be constructed by converting (2.1) to a Volterra integral equation and solving by successive approximations. For example, the wave function normalized at $-\infty$ satisfies the integral equation

$$\phi_0 = 1 + \int_{-\infty}^{x} e^{(x-t)\xi J} Q\phi_0 \, e^{-(x-t)\xi J} \, dt.$$

For general Σ, the wave functions that live on Σ can no longer be obtained by solving a Volterra integral equation. To construct $\eta_\sigma(x, \xi)$ for $\xi \in \Sigma_\sigma$, we proceed as in § 2, equation (2.3): change ν to σ and z to ξ; then repeat the argument exactly as before. This gives a solution of (2.1) on $\Sigma_\sigma \times (-\infty, a]$ that tends to 1 as $x \to -\infty$.

Similarly, we construct a solution of (2.1) on Σ_σ that tends to 1 at $+\infty$ and denote it by ρ_σ. As in the regions Ω_ν we may connect these two solutions by a matrix, say $S_\sigma : \eta_\sigma = \rho_\sigma S_\sigma^x$. Let us define the *Jost* functions of (2.1) as those solutions that live on Σ and are bounded for all x. To obtain them, we must perform a factorization of S_σ. We try to factor $S_\sigma = (1 + U_\sigma)\delta_\sigma(1 + L_\sigma)^{-1}$, where $(\Pi_0^\xi + \Pi_-^\xi)U_\sigma = 0$ and $\Pi_+^\xi L_\sigma = 0$. If this factorization can be performed, then the Jost function is given by

$$m_\sigma = \eta_\sigma(1 + L_\sigma)^x = \rho_\sigma(1 + U_\sigma)^x \delta_\sigma.$$

From the triangular properties of L_σ and U_σ, it is easily seen that $m_\sigma(x, \xi)$ is bounded for all x and tends to 1 as $x \to -\infty$.

Presumably, those values of ξ for which the factorization cannot be performed correspond to poles embedded in the continuous spectrum, but I have not seen how to treat this case. We may call the matrices $\{S_\sigma\}$ the *Jost* scattering data. In the case $J^* = -J$ there is a transformation from the Jost scattering data to the Riemann-Hilbert data. This transformation is worked out for the general case in the Appendix to this paper.

We have thus constructed a global frame on E that tends to the standard frame as $|x|$ and z tend to infinity and in which D_x has the form $d/dx - zJ$.

We have thus replaced the scattering data by a set of transition matrices for a sectionally holomorphic frame. This viewpoint brings the scattering problem in line with the treatment of the self-dual Yang-Mills equation, in which the transition matrices of a holomorphic vector bundle play the role of the scattering data. When considering the associated nonlinear evolution equations with D_x as an isospectral problem, the transition matrices evolve linearly with the flow.

Now let us return to the case when it is not possible to construct a frame with only two regions of the real line $(-\infty, 0)$ and $(0, \infty)$. It is always possible to decompose the line into a finite number of subintervals on which Fredholm equations such as (2.3) have solutions that are analytic in Ω_ν and tend to 1 as $z \to \infty$. In that case we have a set of transition matrices $N_{\sigma,j}$, $R_{\sigma,j}$ and $S_{\nu,j}$ that satisfy the compatibility equations (3.2) for each j. Thus in the general case we must cover the base space with a larger number of open sets. We must correspondingly take a larger set of transition matrices for our data.

Instead of choosing more subintervals of the real line R, Zhou [16] breaks $P_1(C)$ into more regions, as indicated in Fig. 3.1. Outside the large circle he uses the wave

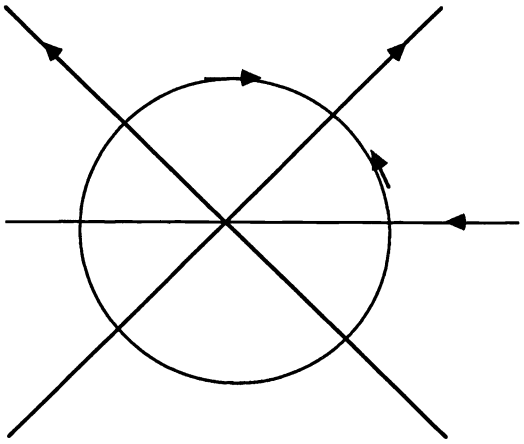


FIG. 3.1. *Support of scattering data in Zhou's method.*

function m constructed by Beals and Coifman, which is regular for sufficiently large z. Inside the circle, he constructs, in each sector, analytic solutions that are normalized at $-\infty$ or at $+\infty$. Since the regions inside the circle are bounded, he does not have to contend with the difficulties in the asymptotic behavior of these solutions as $z \to \infty$.

In any event, to construct a sectionally holomorphic frame for the bundle, we are forced to cover the base space B with a certain minimal number of open sets, either partitioning the real line, or partitioning $P_1(C)$.

The transition matrices are not the same as the scattering data. The scattering data is the minimal data needed to reconstruct the potential Q. The first step in the solution of the inverse scattering problem is to reconstruct the transition matrices from the scattering data.

However, from the point of view of the integration of the nonlinear evolution equations, it is sufficient to work with the transition matrices, since they also evolve linearly with time. The differential equation (2.1) arises as an isospectral problem for a commuting hierarchy of Hamiltonian flows (cf. [2], [3], [8], [11]). These flows arise as "zero-curvature" conditions for a flat connection whose components are

$$D_x = \frac{\partial}{\partial x} - u, \quad D_{t_n} = \frac{\partial}{\partial t_n} - v_n, \quad D_{z^*} = \frac{\partial}{\partial z^*},$$

where $u = zJ + Q$ and $v_n = z^n K + B_n(z, Q)$ (K diagonal) where B_n is a polynomial in Q and its derivatives. The zero-curvature condition $[D_x, D_t] = [\partial/\partial x - u, \partial/\partial t_n - v_n] = 0$ leads to a hierarchy of nonlinear evolution equations in Q.

The linear evolution of the scattering data can also be obtained from a "zero-curvature" condition. Under the sectionally holomorphic gauge transformation g the components of this connection are transformed into

$$\frac{\partial}{\partial x} - zJ, \quad \frac{\partial}{\partial t_n} - z^n K, \quad \frac{\partial}{\partial z^*} + \mu,$$

respectively. Since the flatness of the connection is preserved under the gauge transformation,

$$\left[\partial^* + \mu, \frac{\partial}{\partial x} - zJ\right] = 0, \qquad \left[\partial^* + \mu, \frac{\partial}{\partial t_n} - z^n K\right] = 0.$$

These zero-curvature conditions lead, at least formally, to *linear* evolution equations on the "scattering data" μ:

$$\frac{\partial \mu}{\partial x} = z[J, \mu], \qquad \frac{\partial \mu}{\partial t_n} = z^n[K, \mu].$$

It is useful to introduce an orientation of Σ as indicated in Fig. 3.2.

The orientation of the rays induces an orientation of the complement $\Omega = C \backslash \Sigma$, and Ω can be written as the union of two domains, Ω_+ and Ω_-. Letting $m_\pm$, $\eta_\pm$, etc., denote the sections on $\Omega_\pm$, we can write the transitions more compactly as follows:

$$\eta_+ = \eta_- N^x, \quad \eta_\pm = \rho_\pm S_\pm^x, \quad \rho_+ = \rho_- R^x.$$

4. Winding number constraints. The continuous and discrete components of the scattering data (viz. the poles and principal factors at the poles and the transition matrices V_σ) are not independent, but must satisfy certain "winding number constraints." These were introduced by Bar-Yaacov [1] and play a fundamental role in the recovery of the transition matrices from the scattering data.

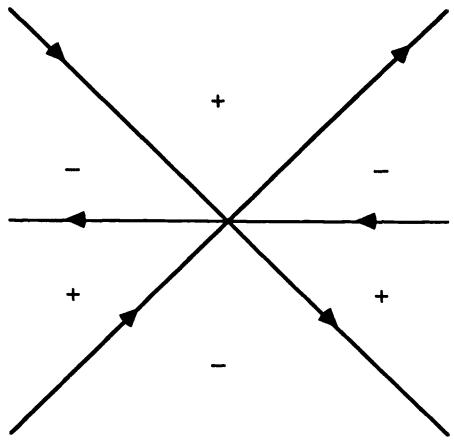


FIG. 3.2. *Oriented system of rays.*

Remark. In [5], which is concerned primarily with first-order systems obtained from a class of nth-order scalar problems, the winding number constraints are satisfied automatically. This is because the first-order systems derived from the scalar case inherit a special symmetry.

The winding number constraints can be derived as a consequence of (3.2). We present the proof here, and also give a more explicit description of the winding number constraints in the case of general scattering data.

The case where all the eigenvalues of J are purely imaginary is somewhat simpler to explain, and we will treat that case first. Let us write $J = i \operatorname{diag}(\lambda_1, \lambda_2 \cdots \lambda_n)$, where the λ_j are real and $\lambda_1 > \lambda_2 > \cdots > \lambda_n$. In this case there are only two sectors, the upper and lower half planes, which we denote by Ω_+ and Ω_-. The permutation σ associated with the rays separating $\Omega_\pm$ is the product of transpositions: $\sigma = (1, n)(2, n-1) \cdots$; it simply reverses the order of the integers $1, \cdots, n$. The permutations $\nu_\pm$ associated with $\Omega_\pm$ are $\nu_+ =$ identity, and $\nu_- = \sigma$.

Let N, R, and $S_\pm$ denote the transition matrices between the four patches $\rho_\pm$ and $\eta_\pm$. For $x = 0$ we have $\eta_+ = \rho_+ S_+$, $\rho_+ = \rho_- R$, $\eta_- = \rho_- S_-$, and $\eta_+ = \eta_- N$; hence

$$S_- N = R S_+. \tag{4.1}$$

Regarding the transition matrices as mappings between the bundle patches $\rho_\pm$, $\eta_\pm$, we may represent these relations schematically as a commuting diagram, as in Fig. 4.1. Let the triangular factorizations of $S_\pm$ be

$$S_+ = (1+U_+)\delta_+(1+L_+)^{-1}, \qquad S_- = (1+L_-)\delta_-(1+U_-)^{-1}, \tag{4.2}$$

where $L_\pm$ and $U_\pm$ are, respectively, lower and upper triangular matrices in the $\nu_\pm$ representations. These factorizations are consistent with the asymptotic conditions satisfied by $\eta_\pm$ and $\rho_\pm$. Now (4.1) is

$$(1+L_-)\delta_-(1+U_-)^{-1}N = R(1+U_+)\delta_+(1+L_+)^{-1};$$

and this may be written as

$$V = (1+U_-)^{-1}N(1+L_+) = \delta_-^{-1}(1+L_-)^{-1}R(1+U_+)\delta_+, \tag{4.3}$$

where $V(\xi)$ is the jump of m across the real axis ($m_+(x,\xi) = m_-(x,\xi)V(\xi)^x$).

To derive the winding number constraints on the scattering data, we need the following two lemmas from the theory of matrices.

LEMMA 4.1. *Let L and U denote general lower and upper triangular matrices, and A an arbitrary matrix. Let $d_k^\pm$ denote the upper and lower $k\times k$ minors. Then*

(i) $d_k^+ LA = (d_k^+ L)(d_k^+ A)$,
(ii) $d_k^+ AU = (d_k^+ A)(d_k^+ U)$,
(iii) $d_k^- AL = (d_k^- A)(d_k^- L)$,
(iv) $d_k^- UA = (d_k^- U)(d_k^- A)$.

LEMMA 4.2. *A matrix T has the triangular factorization $T = (1+L)\delta(1+U)$ if and only if none of its upper minors vanish. Furthermore, its upper minors are given by $d_k^+(T) = d_k^+\delta = \delta_1\delta_2\cdots\delta_k$, where $\delta = \operatorname{diag}(\delta_1, \delta_2, \cdots, \delta_n)$. Similarly, T has the factorization $T = (1+U)\delta(1+L)$ if and only if none of its lower minors vanish, and in that case $d_k^-(T) = d_k^-\delta = \delta_{n-k+1}\cdots\delta_n$.*

The transition matrices have a certain structure, namely, Theorem 4.3.

THEOREM 4.3. *The matrices N and R have (in the standard, identity representation) the triangular factorizations $N = (1+\text{upper})(1+\text{lower})$, $R = (1+\text{lower})(1+\text{upper})$, hence $d_k^- N = 1 = d_k^+ R$.*

Proof. Following [2], approximate Q by potentials of compact support; then $\eta_\pm(x,z)$ are entire functions of z and N is also an entire function of z. We have $N(z)^x = \eta_-(x,z)^{-1}\eta_+(x,z)$. For $x < \min \operatorname{supp} Q$ the wave functions $\eta_\pm$ satisfy the free wave equation $D_z\eta = 0$; hence for $x \ll 0$,

$$\eta_\pm(x,z) = A_\pm(z)^x = e^{xzJ}A_\pm(z)\,e^{-xzJ}.$$

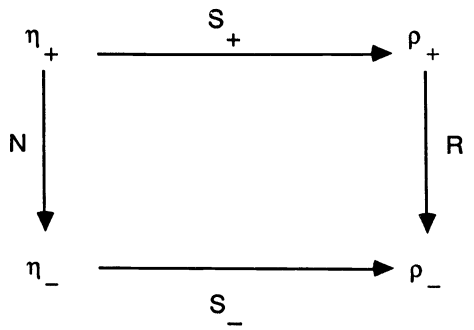


FIG. 4.1. *Diagram of the transition matrices for $J^* = -J$.*

In order that $\eta_\pm$ satisfy the asymptotic conditions as $z \to \infty$ and as $x \to -\infty$, we must have

$$A_+ = 1 + \text{lower} \quad \text{and} \quad A_- = 1 + \text{upper},$$

hence the result holds when Q has compact support. The general case now follows by approximating a general potential by potentials with compact support. The same proof also holds at $+\infty$.

THEOREM 4.4. *The lower minors of V are $1: d_k^-(V) = 1$ for each k. The upper minors satisfy the equation $d_k^+(\delta_+) = (d_k^+\delta_-)(d_k^+ V)$ on the line* $\operatorname{Im} z = 0$. *This may be restated as a Riemann-Hilbert problem for the diagonal matrix δ:*

$$\delta_k^+ = \delta_k^- \phi_k(\xi) \tag{4.4}$$

where

$$\phi_k = \frac{d_k^+ V}{d_{k-1}^+ V}$$

where δ_k is the kth entry of δ, and $d_0^+ V = 1$.

Proof. We have $d_k^- V = d_k^-(1+U_-)^{-1} N(1+L_+) = d_k^- N = 1$ from (2.3), Lemmas 4.1 and 4.2, and Theorem 4.3. Similarly,

$$1 = d_k^+ R = d_k^+(1+L_-)\delta_-(V)\delta_+^{-1}(1+U_+)^{-1} = d_k^+(\delta_-) d_k^+ V d_k^+ \delta_+^{-1}.$$

Equation (4.4) follows directly from this result and the relation $d_k^+ \delta = \delta_1 \delta_2 \cdots \delta_k$. □

The winding number constraints are an immediate consequence of the Riemann-Hilbert problem for $\delta^\pm$. Let $f^\pm$ be any function analytic in $\Omega_\pm$ (upper and lower half planes), tending to 1 as $z \to \infty$, and which does not vanish on the real axis. Let $N(f^\pm) = Z(f^\pm) - P(f^\pm)$, where Z and P are, respectively, the numbers of zeros and poles of $f^\pm$ in $\Omega_\pm$, counted according to multiplicity. Assume both Z and P are finite. From (4.4) we have $d \arg \delta_k^+ = d \arg \delta_k^- + d \arg \phi_k$; integrating this identity along the real line and applying Rouché's theorem, we get

$$N(\delta_k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d \arg \phi_k(\xi) \tag{4.5}$$

where $N(\delta_k)$ is the total number of zeros minus poles of the function δ_k in $\Omega = \Omega_+ \cup \Omega_-$.

We have thus proved Theorem 4.5.

THEOREM 4.5. *The Riemann-Hilbert problem* (4.4) *and the winding number constraints* (4.5) *follow from* (4.1), *the condition that the diagram of the transition matrices commute.*

A method of determining the integers $N(\delta_k)$ from the principal factors of the scattering data is given below in Theorem 4.8.

We now turn to the general case in which the matrix J has complex eigenvalues. We first discuss the structure of the transition matrices. The permutation σ has a very specific form [17]. For $z \in \Omega_\nu$, $\operatorname{Re} z\lambda_{\nu(j)}$ is nondecreasing. Therefore for $\xi \in \Sigma_\sigma$, both $\operatorname{Re} \xi\lambda_{\nu(j)}$ and $\operatorname{Re} \xi\lambda_{\mu(j)}$ are nondecreasing in j. Let us work in the ν representation (i.e., choose a representation in which $\nu =$ identity) so that $\operatorname{Re} \xi\lambda_j$ is nondecreasing in j. There are integers $\alpha_1 < \alpha_2 < \cdots < \alpha_s$ such that σ leaves each $I_p = \{j: \operatorname{Re}(\xi\lambda_j) = \alpha_p\}$ invariant. Moreover, the I_p are consecutive sets of integers: $I_p = \{n_{p-1}+1, \cdots, n_p\}$, where $1 \leqq n_1 < n_2 \cdots n_s \leqq n$; and σ reverses the order in each I_p. Thus σ is the product of transpositions: $\sigma = (1, n_1)(2, (n_1 - 1)) \cdots (n_s, n)((n_s + 1), (n-1))$. For example, the situation for the 3×3 and 4×4 cases, where the elements of J are, respectively, the third and fourth roots of unity, are depicted in Figs. 4.2 and 4.3.

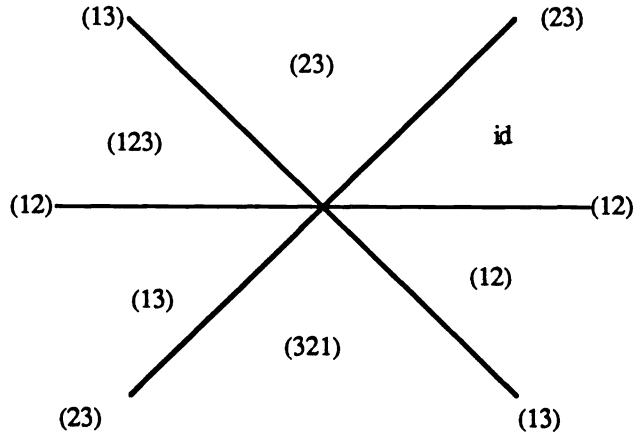


FIG. 4.2. *Sectors for the diagonal matrix* $J = \text{diag}(\omega, \omega^2, 1)$.

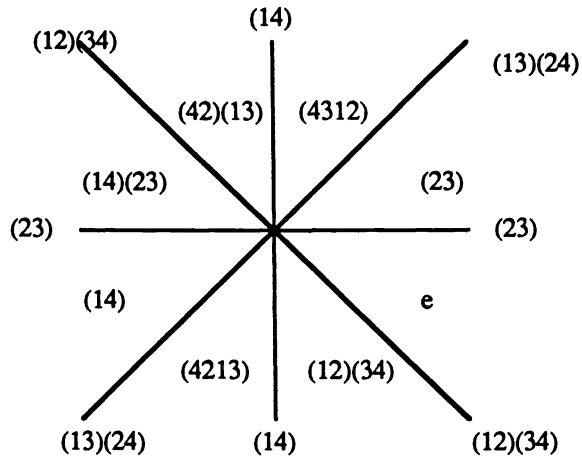


FIG. 4.3. *Sectors for* $J = \text{diag}(-1, -i, i, 1)$.

THEOREM 4.6. *The transition matrix* V_σ *that connects* m_ν *and* m_μ *across* Σ_σ *is block diagonal:* $V_{ij} = 0$ *if* $\text{Re}\,\xi\lambda_i \neq \text{Re}\,\xi\lambda_j$. *Its diagonal blocks factor as* $(1 + upper)(1 + lower)$.

Each transition matrix N_σ *is block lower triangular in the* ν *representation with diagonal blocks of the form* $(1 + upper)(1 + lower)$; *each* R_σ *is block upper triangular in the* ν *representation with diagonal blocks of the form* $(1 + lower)(1 + upper)$.

The transition matrices S_ν *factor as* $(1 + U_\nu)(\delta_\nu)(1 + L_\nu)^{-1}$ *where* U_ν *is upper and* L_ν *is lower triangular in the* ν *representation. In the* ν *representation (with* Ω_ν *and* Ω_μ *as in Fig.* 1.1) L_μ *is block lower triangular with upper triangular diagonal blocks, whereas* U_μ *is block upper triangular with lower triangular diagonal blocks.*

Finally,

$$V_\sigma = \delta_\mu^{-1}(1 + U_\mu)^{-1} R_\sigma (1 + U_\nu)\delta_\nu = (1 + L_\mu)^{-1} N_\sigma (1 + L_\nu). \tag{4.6}$$

Proof (cf. [2]). For $\xi \in \Sigma_\sigma$ we have $(m_\mu^{-1} m_\nu)_{ij} = (V^x)_{ij} = (e^{x\xi J} V e^{-x\xi J})_{ij} = V_{ij}(\xi) \exp\{x\xi(\lambda_i - \lambda_j)\}$. Since m_ν is uniformly bounded on $-\infty < x < \infty$, we must have $V_{ij} = 0$ whenever $\text{Re}\,\xi(\lambda_i - \lambda_j) \neq 0$. Thus, V_σ can be written as a sum of block diagonal matrices $(\oplus_p V_p)$, $1 \leqq p \leqq s$.

For potentials of compact support, $\eta_\nu(x,z) = e^{xzJ}(A_\nu(z))\, e^{-xzJ}$ for $x < \text{supp } Q$. Since $\eta_\nu \to 1$ as $x \to -\infty$, A_ν must be lower triangular in the ν representation. The matrix A_μ is converted to the ν representation by conjugating with σ^{-1}: $A'_\mu = \sigma^{-1}A_\mu\sigma = A_\mu T_\sigma$. From the particular form of σ it is easily seen that A'_μ is block lower triangular with its diagonal blocks upper triangular. Thus in the case of compactly supported potentials, $N_\sigma = (A'_\mu)^{-1}A_\nu$ has the form given in the statement of the theorem. The general result follows by approximating Q by potentials of compact support. The proof of the case for the R_σ is the same.

We know that $S_\mu = (1+U_\mu)(\delta_\mu)(1+L_\mu)^{-1}$ where U_μ and L_μ are, respectively, upper and lower triangular in the μ representation. Their form in the ν representation is given by $U_\mu T_\sigma$ and $L_\mu T_\sigma$, respectively, and the conclusion of the theorem is obtained without difficulty.

Equation (1.6) may be written as follows:

$$\delta_\mu^{-1}(1+U_\mu)^{-1}R_\sigma(1+U_\nu)\delta_\nu = (1+L_\mu)^{-1}N_\sigma(1+L_\nu).$$

From this and the relations $m_\nu = \eta_\nu(1+L_\nu)^x = \rho_\nu(1+U_\nu)^x\delta_\nu$, etc., we obtain (2.6). This completes the proof of Theorem 4.6.

We now turn to the extension of Theorem 4.4 to the sector case. The previous arguments can be repeated, with appropriate modifications for taking into account the block triangular nature of the matrices involved. Taking lower $k\times k$ minors of V_σ, we have $d_k^- V_\sigma = d_k^-[(1+L_\mu)^{-1}N_\sigma(1+L_\nu)] = d_k^-[(1+L_\mu)^{-1}N_\sigma]$ by Lemma 4.1. This time N_σ and $(1+L_\mu)^{-1}$ are block lower triangular; so the lower $k\times k$ minors may be computed by taking determinants of the diagonal blocks. But the diagonal blocks of $(1+L_\mu)^{-1}$ are upper triangular, and the diagonal blocks of N_σ are $(1+\text{upper})\times(1+\text{lower})$. So $d_k^-[(1+L_\mu)^{-1}N_\sigma] = d_k^{-1}V_\sigma = 1$.

Now taking upper $k\times k$ minors of (4.5) we get, by the same arguments, $1 = d_k^+R_\sigma = d_k^+(1+U_\mu)^{-1}R_\sigma(1+U_\nu) = d_k^+(\delta_\nu^{-1}V_\sigma\delta_\mu)$ in the ν representation. Explicitly, we have

$$\begin{aligned} 1 &= d_k^+(\nu^{-1}\delta_\nu^{-1}\nu\nu^{-1}V_\sigma\nu\nu^{-1}\delta_\mu\nu) \\ &= d_k^+(\nu^{-1}\delta_\nu^{-1}\nu)d_k^+(\nu^{-1}V_\sigma\nu)d_k^+(\nu^{-1}\delta_\mu\nu). \end{aligned}$$

We have proved Theorem 4.7.

THEOREM 4.7. *The lower minors of V_σ are all* 1 *along Σ_σ and the upper minors satisfy*

$$\delta_k^\nu = \delta_k^\mu \phi_{k,\sigma}(\xi), \qquad \xi\in\Sigma_\sigma, \tag{4.7}$$

where

$$\phi_{\nu(k)}(\xi) = \frac{d_k^+(\nu^{-1}V_\sigma\nu)}{d_{k-1}^+(\nu^{-1}V_\sigma\nu)}, \qquad \xi\in\Sigma_\sigma \tag{4.8}$$

(*again, we are taking* $d_0(V) = 1$).

Taking arguments and then the differential of (4.7), we get

$$d\arg\delta_k^\nu - d\arg\delta_k^\mu = d\arg\phi_k(\xi), \qquad \xi\in\Sigma_\sigma, \quad \sigma = \mu^{-1}\nu.$$

Integrating this identity over the rays Σ_σ, we get

$$N(\delta_k) = \frac{1}{2\pi}\int_\Sigma d\arg\phi_k(\xi), \tag{4.9}$$

where $\Sigma = \cup_\sigma \Sigma_\sigma$ and $N(\delta_k)$ is the total index of the function δ_k in the complex plane:

$$N(\delta_k) = \frac{1}{2\pi}\int_\Sigma d\arg\frac{\delta_k^+(\xi)}{\delta_k^-(\xi)}.$$

In this integral the functions $\delta_k^+(\xi)$ and $\delta_k^-(\xi)$ denote the limiting values of δ_k on Σ_σ from the regions Ω_ν and Ω_μ, respectively.

By Rouché's theorem, $N(\delta_k)$ is the number of zeros minus the number of poles of δ_k in Ω. These in turn can be related in a precise way to the locations of the poles in the principal factors L_ν^s of the scattering data as follows. First note that

$$\delta_{jj}^\nu = \frac{\Delta_{n-j+1}^-(z)}{\Delta_{n-j}^-(z)}, \qquad z \in \Omega_\nu,$$

where $\Delta_j^-(z) = d_j^-(S_\nu)$. For a function f meromorphic in Ω let $Z_\nu(f)$ and $P_\nu(f)$ denote, respectively, the number of zeros and poles of f in Ω_ν, and let $N_\nu = Z_\nu - P_\nu$. Then $N_\nu(\delta_{jj}) = Z_\nu(\Delta_{n-j+1}^-) - Z_\nu(\Delta_{n-j}^-)$.

Carrying out the factorization $S_\nu(1+L_\nu) = (1+U_\nu)\delta_\nu$, we find that the lower minor Δ_{n-j}^- appears in the denominator of the jth column of L, for $j = 1 \cdots n-1$. So if P_j^ν is the total order of the poles in the jth column of L_ν, in the ν representation, then $P_j^\nu = Z(d_{n-j}^-(S_\nu))$, and $N_\nu(\delta_{ij}) = P_{j-1}^\nu - P_j^\nu$. (We set $P_0^\nu = Z_\nu(d_n^-(S_\nu)) = 0$.) The total index, or degree, of δ_{jj} in Ω is obtained by summing over ν. Thus the winding numbers of the diagonal elements can be obtained directly from the principal factors L_ν^s.

Choose the arguments of $\phi_k(\xi)$ to tend to zero at infinity. Then

$$\sum_\sigma \int_{\Sigma_\sigma} d \arg \phi_{k,\sigma}(\xi) = -\sum_\sigma \arg \phi_{k,\sigma}(0).$$

We have proved Theorem 4.8.

THEOREM 4.8. *The winding number constraints on the scattering data are*

$$P_{k-1} - P_k = -\frac{1}{2\pi} \sum_\sigma \arg \phi_{k,\sigma}(0), \tag{4.10}$$

where the functions ϕ_k are defined by (4.8), *their arguments are chosen to tend to zero as $\xi \to \infty$ along the ray Σ_σ, and*

$$P_k = \sum_\nu P_k^\nu$$

with P_k^ν equal to the number of poles in the kth column of L_ν in the ν representation.

5. Reconstruction of the transition matrices from the scattering data. The inverse problem is to reconstruct the potential Q given scattering data s.d. $= \{V(\xi), z_j, L_\nu\}$ that satisfy the necessary constraints $d_k^-(V) = 1$, $(V-1) \in S(\Sigma)$, and the winding number constraints (4.10). This is done in [2], [5] for the case where all the poles of m are simple, and the principal factors L_ν^s at a pole $z_j \in \Omega_\nu$ have the form (in the ν representation)

$$L = \frac{1}{z - z_j} cE_{k,k+1},$$

where $E_{k,k+1}$ is the matrix with a 1 in the $(k, k+1)$ entry and zeros everywhere else, and c is a constant. Such scattering data is "generic" in the sense that it is the scattering data for a dense class of potentials in L_1, but this class is not invariant under Bäcklund transformations [12], [17].

THEOREM 5.1. *The transition matrices S, R, and N can be uniquely reconstructed from scattering data* s.d. $= \{V(\xi), z_j, L_\nu(z)\}$ *for which $d_k^-(V) = 1$, $(V-1) \in S(\Sigma)$, and for which the winding number constraints* (4.10) *are satisfied.*

Proof. Given V and the factors $1+L_\nu$, N is obtained immediately from the formula $N_\sigma = (1+L_\mu)V_\sigma(1+L_\nu)^{-1}$. The diagonal factors δ_ν of S_ν are obtained by solving the (scalar) Riemann-Hilbert problems (4.7) given the location and multiplicity of the zeros and poles of δ_ν. We saw how to determine the degree of δ_{jj} from the scattering data. (Below we give an explicit solution to the Riemann-Hilbert problem (4.7).)

We next turn to the construction of the factors $1+U_\nu$ and the transition matrices R_σ. In proving Theorem 4.7, we have shown that $d_k^+ V_\sigma = (d_k^+\delta_\nu)(d_k^+\delta_\mu^{-1})$ in the ν representation. By Lemma 4.2, V_σ has the factorization $V_\sigma = (1+B_\mu)^{-1}\delta_\mu^{-1}\delta_\nu(1+B_\nu)$, where B_ν is strictly upper triangular in the ν representation. This factorization is a purely algebraic problem. Performing these factorizations on each of the rays, we obtain matrices B_ν on $\partial\Omega_\nu$, each of which is strictly upper triangular in the ν representation. Next, using Lemma 5.2 below we construct analytic splittings

$$(1+B_\nu) = (1+R_\nu^0)(1+U_\nu^0), \tag{5.1}$$

where U_ν^0 is upper triangular and analytic in Ω_ν, R_ν^0 is upper triangular and analytic in Ω_ν^c, etc.

Then

$$\begin{aligned} V_\sigma &= (1+U_\mu^0)^{-1}(1+R_\mu^0)^{-1}\delta_\mu^{-1}\delta_\nu(1+R_\nu^0)(1+U_\nu^0) \\ &= \delta_\mu^{-1}(1+U_\mu)^{-1}(1+R_\mu)^{-1}(1+R_\nu)(1+U_\nu)\delta_\nu \\ &= \delta_\mu^{-1}(1+U_\mu)^{-1}R_\sigma(1+U_\nu)\delta_\nu, \end{aligned}$$

where $1+U_\nu = (1+\delta_\nu U_\nu^0\delta_\nu^{-1})$, $1+R_\nu = (1+\delta_\nu R_\nu^0\delta_\nu^{-1})$, and $R_\sigma = (1+R_\mu)^{-1}(1+R_\nu)$. We have now constructed N_σ, R_σ, S_μ, and S_ν in such a way that (3.2) is valid.

The analytic splittings (5.1) are obtained through an application of Lemma 5.2 below. Let C be the union of two adjacent rays of Σ, with the orientation inherited from Σ, as in Fig. 5.1. Denote by $\Sigma_\pm$ the components of the complex plane lying to the "left" and "right" of C with the given orientation.

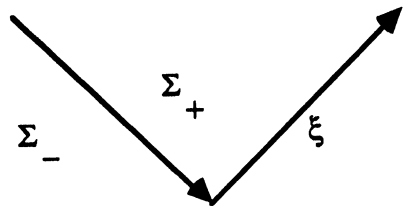


FIG. 5.1

LEMMA 5.2. *Let C and $\Sigma_\pm$ be as shown in Fig. 5.1. Let L be a strictly lower triangular matrix on the contour Σ, with coefficients in $S(\Sigma)$. Then $1+L$ has an analytic splitting $(1+L) = (1+F_+)(1+F_-)$ where $F_\pm$ are the boundary values of matrices analytic in $\Sigma_\pm$, respectively. Similarly, we can factor $(1+L) = (1+F'_-)(1+F'_+)$.*

Proof. Though not Abelian, the group of triangular matrices is solvable, and the factorization problem can be carried out by solving a sequence of scalar problems. Suppose we want to carry out the factorization $(1+F)(1+L) = (1+G)$, where F is analytic in Σ_-, G is analytic in Σ_+, and all matrices are strictly lower triangular.

Let $\Pi_\pm$ denote the projections of functions $f \in S(\Sigma)$ into the boundary values of functions analytic, respectively, in Σ_+ and Σ_-. These projections are constructed in the usual way as follows. Define

$$F(z) = \frac{1}{2\pi i}\int_\Sigma \frac{f(t)}{t-z}\,dt.$$

Then $\Pi_\pm f(\xi) = \lim_{z \to \xi \pm} F(z)$, where the limits are taken as z approaches $\xi \in \Sigma$ from $\Sigma_\pm$, respectively. At points away from the origin, the projections $\Pi_\pm f(\xi)$ are given by the usual Plemelj relations:

$$\Pi_\pm f(\xi) = \pm \frac{f(\xi)}{2} + \frac{P}{\pi i} \int_\Sigma \frac{f(t)}{t - \xi}\, dt.$$

The analysis is more complicated at $z = 0$, but $\Pi_\pm f$ tend to constants as $z \to 0$ from $\Sigma_\pm$, and $(\Pi_+ - \Pi_-) f(0) = f(0)$.

Returning to the proof of the lemma, first choose an F^1 for which the only nonzero terms are immediately below the diagonal: $F^1_{j,j-1} \neq 0$ for $j = 2, \cdots, n$. We use F^1 to annihilate the first subdiagonal of L by setting

$$F^1_{n,n-1} = -\Pi_- L_{n,n-1}, \qquad L^1_{n,n-1} = \Pi_+ L_{n,n-1}.$$

We now have a matrix $(1 + L^1) = (1 + F^1)(1 + L)$ where all terms on the diagonal and first subdiagonal of L^1 vanish. We then repeat the process with the next subdiagonal by choosing an F^2 with nonzero terms on the second subdiagonal, and so forth. The matrix $(1 + F)$ is obtained as the product $(1 + F) = (1 + F^n) \cdots (1 + F^1)$. This proves the lemma.

We apply this lemma to obtain the factorizations (5.1), with $C = \partial\Omega_\nu$, and $\Sigma_+ = \Omega_\nu$.

This essentially completes the proof of Theorem 5.1 except for the solution of the Riemann-Hilbert problem (4.7). An existence proof was given for this problem in [2], using induction on the number of rays Σ_σ. However, a simple constructive proof may be given as follows.

Let $\Sigma = \cup_i \Sigma_j$, where Σ_i are rays extending from the origin, and let ϕ_j be smooth functions in $L_1(\Sigma_j)$, with $\phi_j \to 1$ as $\xi \to \infty$ along the ray. (Recall that the scattering transform preserves the Schwartz class, so that the scattering data will be smooth.) We want to solve the following Riemann-Hilbert problem.

Construct $f(z)$,

(i) Meromorphic in $\Omega = C \backslash \Sigma$; $f(z) \to 1$ as $z \to \infty$ in Ω.

(ii) $f^+(\xi) = f^-(\xi)\phi_j(\xi)$, $\xi \in \Sigma_j$, where $+$ and $-$ denote the limits from the left and right, respectively, Σ_j being oriented from zero to infinity.

(iii) f has the prescribed poles and zeros in Ω.

(iv) f tends to a finite, nonzero limit as $z \to 0$ in Ω, independent of the sector.

THEOREM 5.3. *Let Z and P be, respectively, the numbers of prescribed zeros and poles of f in Ω, counted according to multiplicity. Choose the branch of* $\log \phi_j$ *that tends to zero as $\xi \to \infty$ along the ray Σ_j. Let N be defined by*

$$N = -\sum_j \frac{1}{2\pi i} \int_{\Sigma_j} d(\arg \phi_j) = \frac{1}{2\pi i} \sum_j \arg \phi_j(0). \tag{5.2}$$

Then a unique solution to the Riemann-Hilbert problem exists if and only if $N = Z - P$ and

$$|\phi_1(0)\phi_2(0) \cdots \phi_r(0)| = 1. \tag{5.3}$$

Proof. If there are two solutions f_1 and f_2, we form their ratio. Then all singularities (jumps and poles) cancel out, and their ratio forms a function with removable singularities in the entire plane that tends to 1 as $z \to \infty$, so it is identically one, and the solution is unique.

Necessity. A necessary condition that the problem have a solution is that $\phi_1(0)\phi_2(0) \cdots \phi_r(0) = 1$ [2]. This is seen as follows. Number the sectors and the rays by j, $j = 1, \cdots, r$, and let Σ_j be the ray separating Ω_j from Ω_{j+1}. Across Σ_j we have $f_j^+(\xi) = f_{j+1}^-(\xi)\phi_j(\xi)$. Letting $\xi \to 0$, we get $f_j(0) = f_{j+1}(0)\phi_j(0)$; there is no need to

distinguish limits from the right or left at the origin. Therefore $f_1(0) = f_2(0)\phi_1(0) = f_3(0)f_2(0)\phi_2(0)\phi_1(0) = \cdots = f_1(0)\phi_1(0)\phi_2(0)\cdots\phi_r(0)$, and the result follows from the assumption that f does not vanish at zero. The condition (3.3) follows at once; and moreover $\arg\phi_1(0) + \cdots + \arg\phi_r(0) = 2\pi i J$ for some integer J. We will see in the course of the proof of sufficiency that J must be equal to $Z - P$ for f to have the right asymptotic behavior at infinity.

Sufficiency. It is easily seen that

$$F(z) = \exp\left\{\sum_j \frac{1}{2\pi i}\int_{\Sigma_j} \frac{\log\phi_j(t)}{t-z}\,dt\right\}$$

satisfies the correct jump conditions across the rays, tends to 1 as $z\to\infty$, and has no poles or zeros in Ω. Now

$$\frac{1}{2\pi i}\int_{\Sigma_j} \frac{\log\phi_j(t)}{t-z}\,dt = \frac{1}{2\pi i}[\log\phi_j(0)]\log z + O(1) \quad \text{as } z\to 0,$$

so in the neighborhood of $z=0$, F behaves as does z^K where

$$K = \frac{1}{2\pi i}\log[\phi_1(0)\phi_2(0)\cdots\phi_r(0)].$$

Condition (5.3) is a necessary and sufficient condition for F to be single-valued in an Ω neighborhood of $z=0$. Assuming (5.3) is satisfied, $K = N$, where N is given by (5.2), and F behaves as does z^N near $z=0$. So $z^{-N}F$ is bounded near $z=0$ and does not tend to zero.

Choose

$$R(z) = \prod_{j=1}^{Z}(z-z_j)\Big/\prod_{k=1}^{P}(z-\mu_k)$$

to be a rational function with the prescribed zeros and poles of the problem, counted according to multiplicity. Then $z^{-N}R(z)F(z)$ satisfies all the conditions of the problem in the finite plane and is $O(z^{Z-P-N})$ at infinity. It therefore tends to 1 as $z\to\infty$ if and only if $N = Z - P$. This is the winding number constraint.

To see that the condition $N = Z - P$ is necessary, let f be a solution to the Riemann-Hilbert problem. Then $f \div z^{-N}R(z)F(z)$ is an entire function that behaves as does z^{N-Z+P} as $z\to\infty$. If $N - Z + P < 0$, this ratio, hence f, must vanish identically in the entire plane. If $N - Z + P > 0$, then $f = z^{P-Z}R(z)F(z)$, and f has a zero of order $N - Z + P$ at the origin.

6. Singular integral equations and Plemelj's method. Once the transition matrices N_σ and R_σ have been reconstructed the next step in the solution of the inverse scattering problem is to solve the Riemann-Hilbert problem $\partial^* g = g\mu$, where g is the frame $\{\eta_\nu, \rho_\nu, \eta_\sigma, \rho_\sigma\}$. Since m has poles, it was necessary in [2] and [5] to solve a Riemann-Hilbert problem, or singular integral equation, with poles. In the bundle approach used here, the Riemann-Hilbert problems obtained are regular, with no additional singularities to be dealt with. There is an extensive literature on singular integral equations/Riemann-Hilbert problems, going back to Plemelj. The basic approach is to "regularize" the singular integral equation by reducing it to an integral equation of Fredholm type. We present that method here for the somewhat simpler case $J^* = -J$. Then Σ reduces to the real line in the complex plane.

In the general case, it must be shown that the intersection of the rays of Σ at zero do not introduce singularities at the origin (cf. [2], [5]). It would be interesting to extend the Plemelj approach to this situation.

Thus, let $\Sigma = R$ and let $\Omega_\pm$ denote the upper and lower half planes. Assume we are given the transition matrices N, R, and $S_\pm$, with $RS_+ = S_-N$. Let $\eta_\pm(x, \xi) = \lim_{z \to \xi\pm} \eta(x, z)$ for $\xi \in \Sigma$ and define

$$A_\pm(\xi) = \lim_{x \to -\infty} e^{-x\xi J} \eta_\pm(x, \xi)\, e^{x\xi J}, \qquad \xi \in \Sigma.$$

Then $\eta_\pm(x, \xi) = \eta_0 A_\pm^x$ on Σ, and $N = (A_-)^{-1}A_+$. Similarly, define $B_\pm$ to be the corresponding limits of $\rho_\pm$ as $x \to +\infty$, so that $\rho_\pm = \rho_0 B_\pm^x$ and $R = (B_-)^{-1}B_+$. Across Σ we have $\partial^* g = \eta_+ - \eta_-$ for $x \leqq 0$ and $\partial^* g = \rho_+ - \rho_-$ for $x \geqq 0$. Taking $g = \eta_0$ on $\Sigma^- = \Sigma \times \{x \leqq 0\}$, we find that

$$\mu = g^{-1}\partial^* g = \eta_0^{-1}(\eta_+ - \eta_-) = \eta_0^{-1}(\eta_0 A_+^x - \eta_0 A_-^x) = (A_+ - A_-)^x \quad \text{on } \Sigma.$$

Since $\eta(x, z)$ is analytic in Ω_+ and $\eta - 1 \to 0$ as $z \to \infty$ in Ω_+, we have, for $x \leqq 0$,

$$\eta_+(x, z) - 1 = \frac{1}{2\pi i} \int_\Sigma \frac{\eta_0(x, t)A_+(t)^x - 1}{t - z}\, dt, \qquad z \in \Omega_+$$

and

$$0 = \frac{1}{2\pi i} \int_\Sigma \frac{\eta_0(x, t)A_+(t)^x - 1}{t - z}\, dt, \qquad z \in \Omega_-.$$

Taking the limit of the second integral as $z \to \xi-$ for $\xi \in \Sigma$, we obtain

$$0 = -\frac{1}{2}(\eta_0 A_+(\xi)^x - 1) + \frac{P}{2\pi i} \int_\Sigma \frac{\eta_0 A_+(t)^x - 1}{t - \xi}\, dt, \qquad \xi \in \Sigma.$$

By similar arguments applied to $z \in \Omega_+$, we obtain a second singular integral equation

$$0 = \frac{1}{2}(\eta_0 A_-(\xi)^x - 1) + \frac{P}{2\pi i} \int_\Sigma \frac{\eta_0 A_-(t)^x - 1}{t - \xi}\, dt, \qquad \xi \in \Sigma.$$

Summarizing, we have derived a pair of singular integral equations for η_0 viz.

$$(\eta_0(x, \xi)A_\pm(\xi)^x - 1) = \pm \frac{P}{\pi i} \int_\Sigma \frac{\eta_0(x, t)A_\pm(t)^x - 1}{t - \xi}\, dt. \tag{6.1$_\pm$}$$

Conversely, if η_0 satisfies (6.1$_\pm$), then define η by

$$\eta(x, z) - 1 = \frac{1}{2\pi i} \int_\Sigma \frac{\eta_0(x, t)A_\pm(t)^x - 1}{t - z}\, dt \quad \text{for } z \in \Omega_\pm.$$

Taking the limits of η as $z \to \xi\pm$ and using (6.1$_\pm$), we find that $\eta_\pm(x, \xi) = \eta_0(x, \xi)A_\pm(\xi)^x$, hence $\eta_+ = \eta_- N^x$, where $N = A_-^{-1}A_+$, and our Riemann-Hilbert problem is solved. Thus the singular integral equations (6.1$_\pm$) are fully equivalent to the original Riemann-Hilbert problem.

There is also a second pair of integral equations satisfied by ρ_0, together with the asymptotic conditions

$$\lim_{x \to \infty} e^{-x\xi J} \rho_0\, e^{x\xi J} = 1.$$

Given the two solutions of the singular integral equations η_0 and ρ_0, we obtain solutions $\eta_\pm$ and $\rho_\pm$, analytic in $\Omega_\pm$, which satisfy the Riemann-Hilbert problems $\eta_+ = \eta_- N^x$ and $\rho_+ = \rho_- R^x$ on $\Sigma \times \{x \leqq 0\}$ and $\Sigma \times \{x \geqq 0\}$, respectively, with the additional radiation conditions. Suppose these two Riemann-Hilbert problems have unique solutions under the conditions stated. Putting $x = 0$, multiplying the second equation

above by $S_+(\xi)$, and using the identity $RS_+ = S_-N$, we find $\rho_+S_+ = \rho_-(RS_+) = (\rho_-S_-)N$. Hence $\rho_\pm S_\pm$ satisfy the same Riemann-Hilbert problem as $\eta_\pm$. By uniqueness, $\eta_\pm = \rho_\pm S_\pm$ for $x = 0$ and $\xi \in \Sigma$.

In the general case we would have to solve a sequence of Riemann-Hilbert problems on successive intervals of the real line and patch the solutions together.

The potential Q is recovered as follows. Define $w(x, z) = \eta_\pm(x, z)$ for $x \leqq 0$ and $z \in \Omega_\pm$, and $w = \rho_\pm(x, z)$ for $x \geqq 0$ and $z \in \Omega_\pm$. Consider the expression $Q(x, z) = (D_z w)w^{-1}$, where D_z is the derivation

$$D_z = \frac{\partial}{\partial x} - z \operatorname{ad} J.$$

By the analytic properties of w, Q is analytic in $\Omega_+ \cup \Omega_-$ and is bounded as $z \to \infty$. Its only possible singularities consist of jumps across Σ. But on Σ, say for $x \leqq 0$,

$$\begin{aligned} Q_\pm(x, \xi) &= (D_\xi \eta_\pm)\eta_\pm^{-1} = D_\xi(\eta_0 A_\pm^x)(A_\pm^{-1})^x \eta_0^{-1} \\ &= D_\xi(\eta_0)(A_\pm^x)(A_\pm^{-1})^x \eta_0^{-1} + \eta_0 D_\xi(A_\pm^x)(A_\pm^{-1})^x \eta_0^{-1} \\ &= D_\xi(\eta_0)\eta_0^{-1} \end{aligned}$$

since $D_\xi(A_\pm^x) = 0$. Thus Q has the same limits from both sides of Σ and therefore has no jumps across Σ. So Q is entire in z and bounded as $z \to \infty$, hence Q is independent of z.

Let us now return to the reduction of the singular integral equations to a Fredholm system. First rewrite the pair of equations $(6.1_\pm)$ as

$$\eta_0 - 1 = (A_\pm^{-1})^x - 1 \pm \frac{P}{\pi i}\int_\Sigma \frac{(\eta_0 - 1)A_\pm(t)^x A_\pm^{-1}(\xi)^x}{t - \xi}\, dt \pm \frac{P}{\pi i}\int_\Sigma \frac{A_\pm(t)^x - 1}{t - \xi} A_\pm^{-1}(\xi)^x\, dt,$$

where, more specifically, $A_\pm(\xi)^x = e^{x\xi J} A_\pm(\xi)\, e^{-x\xi J}$, etc. Adding these two equations, we get

$$\eta_0 - 1 = k_1(x, \xi) + \int_\Sigma (\eta_0 - 1)K_1(x, t, \xi)\, dt, \tag{6.2}$$

where

$$\begin{aligned} K_1(x, t, \xi) &= \frac{1}{2\pi i}\,\frac{A_+(t)^x A_+^{-1}(\xi)^x - A_-(t)^x A_-^{-1}(\xi)^x}{t - \xi} \\ &= \frac{1}{2\pi i} A_-(t)^x \frac{N(t)^x - N(\xi)^x}{t - \xi} A_+^{-1}(\xi)^x \end{aligned}$$

and

$$k_1(x, \xi) = \frac{1}{2}\left\{1 + \frac{P}{\pi i}\int_\Sigma \frac{A_+(t)^x - 1}{t - \xi}\, dt\right\} A_+^{-1}(\xi)^x + \frac{1}{2}\left\{1 - \frac{P}{\pi i}\int_\Sigma \frac{A_-(t)^x - 1}{t - \xi}\, dt\right\} A_-^{-1}(\xi)^x - 1.$$

Below we show that (6.2) is, in fact, a Fredholm integral equation; but first we need to investigate whether it is equivalent to the original Riemann-Hilbert problem. Let η_0 satisfy (6.2) and define

$$\psi(x, z) = \frac{1}{2\pi i}\int_\Sigma \frac{\eta_0 A_\pm(t)^x - 1}{t - z}\, dt, \qquad z \in \Omega_{-,+}.$$

The limits of $\psi(x, z)$ as $z \to \xi\pm$ are

$$\psi_\pm(x, \xi) = \pm\frac{1}{2}(\eta_0 A_{-,+}^x - 1) + \frac{P}{2\pi i}\int_\Sigma \frac{\eta_0 A_{-,+}^x - 1}{t - \xi}\, dt.$$

Comparing these expressions with equations $(6.1_{\pm})$, we see that the singular integral equations $(6.1_{\pm})$ for η_0 are equivalent to the conditions $\psi_{\pm}(x, \xi) = 0$. Since ψ is analytic in $z \in \Omega$, and tends to zero as $z \to \infty$, we see that ψ must vanish identically.

It is a simple computation to show that if η_0 satisfies (6.1), then ψ satisfies the Riemann-Hilbert problem $\psi_- = \psi_+ N^x$. (Recall that η satisfies the Riemann-Hilbert problem $\eta_+ = \eta_- N^x$.) The Riemann-Hilbert problem satisfied by ψ is called the *accompanying problem* [10], [14]. Thus (6.2) is equivalent to the singular integral equations $(6.1_{\pm})$ provided that the accompanying problem has no homogeneous solutions.

The accompanying problem can, in turn, be reduced to a system of Fredholm integral equations. Rewriting it as $\psi_-(A_+^{-1})^x = \psi_+(A_-^{-1})^x$, and defining $\psi_0 = \psi_-(A_+^{-1})^x = \psi_+(A_-^{-1})^x$, we find by the same arguments as above that

$$\frac{1}{2\pi i}\int_{\Sigma} \frac{\psi_0 A_{\pm}(t)^x}{t - z}\, dt = 0 \quad \text{for } z \in \Omega_{\pm}.$$

Taking limits as $z \to \xi\pm$, we get

$$\psi_0 A_{\pm}(\xi)^x \pm \frac{P}{\pi i}\int_{\Sigma} \frac{\psi_0 A_{\pm}(t)^x}{t - \xi}\, dt = 0.$$

Multiplying these two equations on the right by $A_{\pm}(\xi)^x$, respectively, and adding, we get the homogeneous integral equation

$$\psi_0(x, \xi) + \frac{1}{2\pi i}\int_{\Sigma} \psi_0(x, t)\, \frac{A_+(t)^x (A_+(\xi)^{-1})^x - A_-(t)^x (A_-(\xi)^{-1})^x}{t - \xi}\, dt = 0 \tag{6.3}$$

corresponding to the accompanying problem.

Our next task is to demonstrate that (6.2), (6.3) are indeed Fredholm integral equations on $L_2(\Sigma)$. We do this by showing that $K_1(x, t, \xi)$ is in $L_2(\Sigma \times \Sigma)$. It is clear that if the scattering data is at least Lipschitz continuous, then K_1 is bounded on the diagonal $t = \xi$. A second difficulty arises because Σ is infinite. We must show that

$$\int_{\Sigma}\int_{\Sigma} |K_1(x, t, \xi)|^2\, dt\, d\xi < +\infty,$$

where, for $|K_1(x, t, \xi)|$ we may take the sum of the absolute values of the entries of K_1, or any equivalent norm. We decompose the integral into integrals over the sets $|\xi - t| < 1$ and $|\xi - t| \geqq 1$. Let us consider the integral over $|\xi - t| \geqq 1$ first. Since A_-^x and $(A_+^{-1})^x$ are bounded on Σ (all the exponentials are oscillatory on Σ), we have

$$|K_1(x, t, \xi)| \leqq \text{const.} \left| \frac{N(t)^x - N(\xi)^x}{t - \xi} \right|. \tag{6.4}$$

Now $N(t) = 1 + W(t)$, where W is smooth and decays as $t \to \pm\infty$, so

$$|K_1(x, t, \xi)| \leqq \text{const.} \frac{|W(t)| + |W(\xi)|}{|t - \xi|}.$$

Hence

$$\iint_{|\xi - t| \geqq 1} |K(x, t, \xi)|^2\, dt\, d\xi \leqq \text{const.} \left\{ \iint_{|\xi - t| \geqq 1} \frac{|W(t)|^2 + |W(t)||W(\xi)|}{|\xi - t|^2}\, dt\, d\xi \right\}.$$

The first integral is dominated by

$$\int_{-\infty}^{\infty} |W(t)|^2 \left\{ \int_{|\xi - t| \geqq 1} \frac{d\xi}{|\xi - t|^2} \right\} dt \leqq 2 \int_{-\infty}^{\infty} |W(t)|^2\, dt < +\infty$$

and the second term is dominated by

$$\iint |W(t)||W(\xi)|\, dt\, d\xi = 2\|W\|_1^2.$$

The integral over $|\xi - t| < 1$ is bounded by

$$\iint_{|s-t|<1} \left|\frac{N(t)^x - N(s)^x}{t-s}\right|^2 ds\, dt = \int_{-\infty}^{\infty} \left\{ \int_{s-1}^{s+1} \left|\frac{N(t)^x - N(s)^x}{t-s}\right|^2 dt \right\} ds. \tag{6.5}$$

Now

$$\frac{N(t)^x - N(s)^x}{t-s} = \frac{1}{t-s}\int_s^t \frac{d}{dy} N(y)^x\, dy = \frac{1}{t-s}\int_s^t e^{xyJ}(N'(y) + x[J, N])\, e^{-xyJ}\, dy,$$

so (6.5) is finite if $g(s) \in L_2(R)$, where

$$g(s) = \sup_{s-1 \leqq y \leqq s+1} |N'(y) + x[J, N(y)]|.$$

Now we are looking for solutions satisfying the asymptotic condition

$$\lim_{x \to -\infty} e^{-x\xi J} \eta_0(x, \xi)\, e^{x\xi J} = 1$$

and a similar condition on ρ_0 at $+\infty$. Multiplying on the left by $e^{-x\xi J}$ and on the right by $e^{x\xi J}$, and putting $w_0(x, \xi) = e^{-x\xi J}\eta_0(x, \xi)\, e^{x\xi J} - 1$, we get the integral equation

$$w_0 = k(x, \xi) + \int_\Sigma e^{x(t-\xi)J} w_0 K(x, t, \xi)\, dt$$

where

$$K(x, t, \xi) = \frac{1}{2\pi i} \frac{A_+(t)\, e^{-x(t-\xi)J} A_+^{-1}(\xi) - A_-(t)\, e^{-x(t-\xi)J} A_-^{-1}(\xi)}{t - \xi}$$

and

$$k(x, \xi) = \frac{1}{2}\left\{1 + \frac{P}{\pi i}\int_\Sigma e^{x(t-\xi)J} \frac{A_+(t) - 1}{t-\xi} e^{-x(t-\xi)J}\, dt\right\} A_+^{-1}(\xi)$$
$$+ \frac{1}{2}\left\{1 - \frac{P}{\pi i}\int_\Sigma e^{x(t-\xi)J} \frac{A_-(t) - 1}{t-\xi} e^{-x(t-\xi)J}\, dt\right\} A_-^{-1}(\xi) - 1.$$

We need to show that the inhomogeneous term $k(x, \xi) \to 0$ as $x \to -\infty$. Since all the exponential terms in $e^{xtJ} A_+(t)\, e^{-xtJ}$ are purely imaginary for $t \in \Sigma_\sigma$, the limit

$$\lim_{x \to -\infty} \frac{P}{\pi i}\int_\Sigma e^{x(t-\xi)J} \frac{A_+(t) - 1}{t-\xi} e^{-x(t-\xi)J}\, dt$$

may be evaluated using the following result from the theory of distributions:

$$\lim_{x \to \pm\infty} \frac{P e^{itx}}{\pi i t} = \pm\delta(t),$$

where P stands for the Cauchy principal value. For the jkth entry we get

$$= -(A_+(\xi) - 1)_{jk} \operatorname{sgn} \operatorname{Im}(\lambda_j - \lambda_k).$$

Now Re $z\lambda_j <$ Re $z\lambda_k$ for $j<k$ and Re $z>0$. Hence sgn Im $(\lambda_j-\lambda_k)>0$ for $j<k$. Since $A_+(t)$ is lower triangular by Theorem 4.6, we get

$$A_+(\xi)-1.$$

Similarly,

$$\lim_{x\to-\infty}\frac{P}{\pi i}\int_\Sigma e^{x(t-\xi)J}\frac{A_-(t)-1}{t-\xi}e^{-x(t-\xi)J}\,dt=-(A_-(\xi)-1)$$

so $k(x,\xi)\to 0$ pointwise as $x\to-\infty$.

With further efforts, $k(x,\xi)$ can be shown to tend to zero in the mean as $x\to\infty$ as well; on the other hand, the dependence of kernel $K(x,\xi)$ on x is the exponential of an imaginary term, so the norm of K remains bounded as $x\to-\infty$. This step requires further analysis.

When $\Sigma=R$ the Riemann-Hilbert problem can also be Fourier transformed to a convolution equation on a semi-infinite line. This approach has been considered in detail by Gokhberg and Krein [9]. In this picture, there is a very striking difference between the 2×2 and $n\times n$ cases. In the $n\times n$ case, the integral operators obtained are of the form (roughly speaking)

$$F\psi(s)=\int_0^\infty f(s-t)\psi(t)\,dt.$$

Such an integral operator F is not even Fredholm. In the 2×2 case (and also the second-order scalar case), the inverse scattering problem can be reduced to the Gel'fand-Levitan-Marchenko integral equation, in which the operator takes the form

$$F\psi(s)=\int_s^\infty f(s+t)\psi(t)\,dt.$$

Here the kernel of the integral operator depends on the *sum* of the arguments; and, under very mild decay conditions on f (which are satisfied in practice) the integral operator is of trace class and the integral equation can be solved by Fredholm's method of determinants and minors. The second- and nth-order cases are thus totally different in character.

7. The "dressing method" for $n\times n$ hierarchies. In this section we develop a parallel of Zakharov and Shabat's dressing method for the Kadomtsev-Petviashvili (KP) hierarchy based on sectionally holomorphic gauge transformations. We first review the dressing method for the KP hierarchy [15]. The KP hierarchy can be obtained by "dressing" the family of differential operators

$$\frac{\partial}{\partial x_n}-D^n,\quad D=\frac{d}{dx},\quad n=2,3,\cdots$$

with upper and lower Volterra integral operators $W_\pm=1+K_\pm$, where

$$K_+\psi(x)=\int_x^\infty K_+(x,y)\psi(y)\,dy$$

and K_- is defined similarly. We obtain, formally, the hierarchy of nth-order differential operators B_n, defined by

$$\left(\frac{\partial}{\partial x_n}-B_n\right)W_\pm=W_\pm\left(\frac{\partial}{\partial x_n}-D^n\right).\tag{7.1}$$

Following Zakharov and Shabat, we will say that the operators $W_\pm$ *dress* the constant coefficient bare operators on the right to the perturbed operators on the left.

Assuming one of the operators, say W_+, is invertible, the following proposition is easily demonstrated [15].

THEOREM 7.1. *Let* $1+F=W_+^{-1}W_-$. *Then* $W_\pm$ *both dress the bare operators* $\partial/\partial x_n - D^n$ *to the same operator* $\partial/\partial x_n - B_n$ *if and only if*

$$[1+F, \partial/\partial x_n - D^n]=0. \tag{7.2}$$

Moreover, in this case, the B_n *are purely differential operators.*

The commutation relations (7.2) are linear evolution equations for the "scattering operator" F:

$$\frac{\partial F}{\partial x_n}=[F, D^n]. \tag{7.3}$$

The basic intertwining relations (7.1) can be rewritten as the operator identity

$$\frac{\partial W}{\partial x_n}W^{-1}=B_n - WD^nW^{-1}.$$

The left side of this identity is an integral operator, whereas WD^nW^{-1} contains both differential and integral terms. Decomposing this identity into its differential and integral parts, we obtain the well-known relationships

$$B_n=[L^n]_+, \quad \frac{\partial W}{\partial x_n}W^{-1}=-[L^n]_-, \quad L=WDW^{-1}, \tag{7.4}$$

where $[L^n]_\pm$ denote the differential and integral parts of the integrodifferential operator $L^n=WD^nW^{-1}$. From the identities $LW=WD$ and $(\partial/\partial x_n - B_n)W=W(\partial/\partial x_n - D^n)$, and the commutation relation $[\partial/\partial x_n - D^n, D]=0$, we obtain the nonlinear evolution equations

$$\frac{\partial L}{\partial x_n}=[(L^n)_+, L]. \tag{7.5}$$

In an exact analogy with the KP hierarchy, the $n\times n$ hierarchies introduced in § 3 can be obtained by constructing a sectionally holomorphic gauge transformation g such that

$$\begin{aligned} g\left(\frac{\partial}{\partial x}-zJ\right)&=\left(\frac{\partial}{\partial x}-u(x,z)\right)g, \qquad u=zJ+Q(x),\\ g\left(\frac{\partial}{\partial t_n}-z^nK\right)&=\left(\frac{\partial}{\partial t_n}-v_n(x,z)\right)g, \qquad v_n=z^nK+B_n(z,Q). \end{aligned} \tag{7.6}$$

We orient Σ as in Fig. 3.1 and denote the associated components of g in $\Omega_\pm$ by $g_\pm$. In analogy with Theorem 7.1, we have Theorem 7.2.

THEOREM 7.2. *Let* g *be a sectionally holomorphic gauge transformation that tends to* 1 *as* $|x|$ *or* $|z|$ *tends to infinity. Let* $g_\pm$ *denote the limiting values of* g *on* Σ *from* Ω_+ *and* Ω_-. *Then* $g_\pm$ *both dress* $(\partial/\partial x-\xi J)$ *and* $(\partial/\partial t_n-\xi^nK)$ *to the same operators* $\partial/\partial x-u(x,\xi)$, $(\partial/\partial t_n-v_n(\xi,Q))$ *if and only if*

$$\left[F, \frac{\partial}{\partial x}-\xi J\right]=0, \qquad \left[F, \frac{\partial}{\partial t_n}-\xi^nK\right]=0, \tag{7.7}$$

where $F=g_-^{-1}g_+$. *Moreover, in this case, both* u *and* v_n *are polynomials in* z.

Proof. The relations (7.7) may be expressed concisely as $D_n g_-^{-1} g_+ = 0$, where D_n are the derivations

$$D_0 = \frac{\partial}{\partial x} - \xi \operatorname{ad} J, \qquad D_n = \frac{\partial}{\partial t_n} - \xi^n \operatorname{ad} K$$

and the intertwining relations (7.6) can be written concisely as

$$v_n = \frac{\partial g}{\partial t_n} g^{-1} + g z^n K g^{-1}, \qquad u = \frac{\partial g}{\partial x} g^{-1} + g z J g^{-1}.$$

(Equivalently, we may write $B_n = (D_n g) g^{-1}$, where $B_0 = Q(x)$.) We will denote $v_0 = u$ and $t_0 = x$. The proof of Theorem 7.2 is based on the following identity, which is easily derived:

$$v_n^+ - v_n^- = g_-(D_j g_-^{-1} g_+) g_+^{-1}, \qquad \xi \in \Sigma, \tag{7.8}$$

where $v_n^\pm$ are the limits of v_n on Σ from $\Omega_\pm$.

If $g_\pm$ both dress $(\partial/\partial x - \xi J)$ and $(\partial/\partial t_n - \xi^n K)$ to the same operators $(\partial/\partial x - u(x, \xi))$, $(\partial/\partial t_n - v_n(\xi, Q))$, then $v_n^+ = v_n^-$ on Σ, and $D_j g_-^{-1} g_+ = 0$ by (7.8). Conversely, suppose (7.7) are satisfied on Σ for a sectionally holomorphic gauge transformation g that tends to 1 as $|x|$ or z tends to ∞, and define $v_n = (\partial g/\partial t_n) g^{-1} + g z^n K g^{-1}$. Then $v_n^+ = v_n^-$ on Σ by (7.8), so v_n is an entire function of z. Since $g \to 1$ as $z \to \infty$, $v_n = O(z^n)$ for $n \geqq 1$, and $u = O(z)$ as $z \to \infty$. Therefore u and v_n are polynomials of degree 1 and n in z, respectively.

Now g has an asymptotic expansion in inverse powers of z in each sector $\Omega_\nu^\pm$ uniform in x in $(-\infty, a]$ or x in $[a, \infty)$ respectively, for any finite a, hence $G = gKg^{-1}$ also has an asymptotic expansion:

$$gKg^{-1} \sim \sum_{j=0}^{\infty} \frac{G_j}{z^j}.$$

Since $v_n = (\partial g/\partial t_n) g^{-1} + (z^n G)$ is a polynomial in z, $v_n = (z^n G)_+$, where $(z^n G)_+$ the polynomial part of the asymptotic expansion of $z^n G$.

We claim that v_n is a polynomial in Q and its derivatives up to order n. The corresponding statement is proved in [11] for the terms in the asymptotic expansion of mKm^{-1} where m is the meromorphic wave function of Theorem 2.1; namely, $(z^n mKm^{-1})_+$ is a polynomial of order n in z whose coefficients are differential polynomials in Q.

The differential polynomial character of z is unaffected by the transformation from m to g. The relation between m and g is given simply by $m = g(1 + L_\nu)^\xi$ in Ω_ν^- and $m = g(1 + U_\nu)^\xi \delta_\nu$ in Ω_ν^+, where,

$$\xi = xzJ + \sum_{j=1}^{\infty} t_n z^n K$$

and $A^\xi = e^\xi A\, e^{-\xi}$. On the other hand, $v_n = z^n K + B_n(z, Q)$, where $B_n(z, Q) = (D_n g) g^{-1}$, and $(D_n g) g^{-1}$ is invariant under right multiplication of g by an element in the kernel of the derivation D_n. (Note that the ker $D_n = \{A^\xi\}$.)

It is easily verified that G satisfies the following differential equations:

$$\frac{\partial G}{\partial t_n} = [v_n, G], \qquad n = 0, 1, \cdots$$

where, as usual, $t_0 = x$ and $v_0 = u$. From the first of these equations ($n = 0$) we obtain the recursion relation

$$[J, G_{j+1}] = \frac{\partial G_j}{\partial x} - [Q, G_j].$$

It follows that

$$\frac{\partial v_n}{\partial x} = \left(z^n \frac{\partial G}{\partial x} \right)_+ = (z^n [u, G])_+ = (z^n [zJ + Q, G])_+$$
$$= [J, (z^{n+1} G)_+] + [Q, (z^n G)_+],$$

hence that

$$[J, (z^{n+1} G)_+] = \frac{\partial}{\partial x} (z^n G)_+ - [Q, (z^n G)_+].$$

Furthermore, the recursion relation $(z^{n+1} G)_+ = G_{n+1} + z(z^n G)_+$ is easily verified. We therefore have Lemma 7.3.

LEMMA 7.3. *The differential polynomials v_n satisfy the recursion relations*

$$[J, v_{n+1}] = \left(\frac{\partial}{\partial x} - \text{ad } Q \right) v_n, \qquad v_{n+1} - z v_n = G_{n+1}.$$

The Hamiltonian hierarchy of flows follow from the zero-curvature relations

$$\left[\frac{\partial}{\partial t_m} - v_m, \frac{\partial}{\partial t_n} - v_n \right] = 0.$$

From the equation for $m = 0$ we get

$$\frac{\partial u}{\partial t_n} = \frac{\partial v_n}{\partial x} + [(z^n G)_+, u]. \tag{7.9}$$

Since $u = zJ + Q$, these equations can also be written

$$\frac{\partial Q}{\partial t_n} = \frac{\partial v_n}{\partial t} - \frac{\partial v_n}{\partial x} + [v_n, u] = \frac{\partial v_n}{\partial x} - [Q, v_n] - z[J, v_n]$$
$$= [J, v_{n+1} - z v_n] = [J, G_{n+1}] = \left(\frac{\partial}{\partial x} - \text{ad } Q \right) G_n$$
$$= [J, G_{n+1}].$$

Appendix. Transformation of the Jost data to the bundle data. In this Appendix we show how to obtain the transition matrices N_σ and R_σ from the matrices S_σ, S_ν. We first prove the following lemma. We use the notation $\eta^{-x} = e^{-x\xi J} \eta\, e^{x\xi J}$.

LEMMA A1. *The limits*

$$\lim_{x \to -\infty} \Pi_0^\xi \eta_\sigma^{-x}, \quad \lim_{x \to -\infty} \Pi_0^\xi m_\pm^{-x}, \quad \lim_{x \to \infty} \Pi_0^\xi \rho_\sigma^{-x}, \quad \lim_{x \to \infty} \Pi_0^\xi m_\pm^{-x}$$

exist for all $\xi \in \Sigma_\sigma$ for which the corresponding wave function is defined and satisfies the appropriate integral equation.

Proof. Let us prove the statement for η_σ, which satisfies the integral equation (2.3). Operate by $e^{-x\xi J}$ on the left and by $e^{x\xi J}$ on the right and apply the projection Π_0^ξ. This projection knocks out the integral to $+\infty$, and it is immediate that

$$\lim_{x \to -\infty} \Pi_0^\sigma \eta_\sigma^{-x} = \Pi_0^\sigma 1 = 1.$$

The wave function m_ν (which satisfies the differential equation in the interior of Ω_ν) satisfies the integral equation

$$m_\nu = 1 + \int_{-\infty}^{x} (\Pi_-^z)\, e^{(x-y)zJ} Q m_\nu\, e^{-(x-y)zJ}\, dy - \int_x^{\infty} \Pi_+^z\, e^{(x-y)zJ} Q m_\nu\, e^{-(x-y)zJ}\, dy,$$

where Π_+^z projects onto the subspace where Re z ad J is strictly positive for all $z \in \Omega_\nu$. We obtain an integral equation for the boundary values $m_\pm$ by taking the limit as z approaches $\xi\pm$. In the limit as $z \to \xi\pm$, the ranges of the projections $\Pi_\pm^\xi$ each contain terms for which Re $\xi(\lambda_j - \lambda_k) = 0$. Conjugating with $e^{-x\xi J}$ and applying Π_0^ξ, we obtain

$$\lim_{x\to-\infty} \Pi_0^\xi m_+^{-x} = 1 - \int_{-\infty}^{\infty} \Pi_0^\xi (Q m_+)^{-\xi}\, dy = 1 + L_\nu,$$

where the matrix L_ν is in the range of the projection Π_0^ξ in the ν representation. The other limits are established in the same way.

Now we can write $\eta_\nu = \eta_\sigma (A_\nu)^x$, $\eta_\mu = \eta_\sigma (A_\mu)^x$ where

$$A_\nu = \lim_{x\to-\infty} \Pi_0^\xi \eta^{-x}.$$

It follows immediately that $N_\sigma = (A_\mu)^{-1}(A_\nu)$. Similarly, we have $R_\sigma = (B_\mu)^{-1}(B_\nu)$, where

$$\rho_\nu = \rho_\sigma B_\nu^x \quad \text{and} \quad B_\nu = \lim_{x\to\infty} \Pi_\sigma^\xi \rho^{-x}.$$

The matrices A_ν and B_ν can be uniquely determined from S_ν, S_σ as follows. The relations

$$\eta_\nu = \rho_\nu S_\nu^x, \quad \eta_\nu = \eta_\sigma (A_\nu)^x, \quad \rho_\nu = \rho_\sigma (B_\nu)^x, \quad \eta_\sigma = \rho_\sigma S_\sigma^x,$$

imply $S_\sigma(A_\nu) = (B_\nu) S_\nu$. Similarly, $S_\sigma(A_\mu) = (B_\mu) S_\mu$. By the triangularity properties of the A_ν and B_ν we conclude that

$$d_k^+ S_\sigma = d_k^+ S_\nu, \qquad d_k^- S_\sigma = d_k^- S_\mu.$$

Given two matrices S_σ and S_ν with the same upper minors there are uniquely determined lower and upper triangular matrices A_ν and B_ν such that $S_\sigma = B_\nu S_\nu A_\nu^{-1}$. Similarly, there are uniquely determined μ-lower and μ-upper matrices A_μ and B_μ such that $S_\sigma = B_\mu S_\mu A_\mu^{-1}$.

Acknowledgments. I acknowledge helpful conversations with a number of colleagues: Greg Anderson, Dick Beals, Percy Deift, Jacek Szmigelski, Xin Zhou, and Victor Zurkowski. In addition, I thank the referees for their patient reading of the manuscript and constructive suggestions.

REFERENCES

[1] D. BAR-YAACOV, *Analytic properties of scattering and inverse scattering for first order systems*, Ph.D. thesis, Yale University, New Haven, CT, 1985.

[2] R. BEALS AND R. COIFMAN, *Scattering and inverse scattering for first order systems*, Comm. Pure Appl. Math., 37 (1984), pp. 39-90.

[3] ———, *Scattering and inverse scattering for first-order systems*: II, Inverse Problems, 3 (1987), pp. 577-593.

[4] R. BEALS, *The inverse problem for ordinary differential operators on the line*, Amer. J. Math., 107 (1985), pp. 281-366.

[5] R. BEALS, P. DEIFT, AND C. TOMEI, *Direct and Inverse Scattering on the Line*, Math. Surveys Monographs, Vol. 28, American Mathematical Society, Providence, RI, 1988.

[6] P. DEIFT AND E. TRUBOWITZ, *Inverse scattering on the line*, Comm. Pure Appl. Math., 32 (1979), pp. 121-251.

[7] P. DEIFT, C. TOMEI, AND E. TRUBOWITZ, *Inverse scattering and the Boussinesq equation*, Comm. Pure Appl. Math., 35 (1982), pp. 567-628.

[8] L. D. FADDEEV AND L. A. TAKHTAJAN, *Hamiltonian Methods in the Theory of Solitons*, Springer-Verlag, Berlin, New York, 1988.

[9] I. TS. GOKHBERG AND M. G. KREIN, *Systems of integral equations on the half line with kernels which depend on the difference of the arguments*, Uspekhi Mat. Nauk, 13 (1958), pp. 3-72; Amer. Math. Soc. Transl., Vol. 14, Providence, RI, 1960, pp. 217-286.

[10] J. PLEMELJ, *Riemannsche Funktionenscharen mit gegebener Monodromiegruppe*, Monatsh. Math. Phys., 19 (1908), pp. 211-245.

[11] D. H. SATTINGER, *Hamiltonian hierarchies on semi-simple Lie algebras*, Stud. Appl. Math., 72 (1985), pp. 65-86.

[12] D. H. SATTINGER AND V. D. ZURKOWSKI, *Gauge theory of Bäcklund transformations*, II, Phys. D, 26 (1987), pp. 225-250.

[13] A. B. SHABAT, *An inverse scattering problem*, J. Differential Equations, 15 (1980), pp. 1299-1307.

[14] N. P. VEKUA, *Systems of Singular Integral Equations*, P. Noordhoff, Groningen, the Netherlands, 1967.

[15] V. E. ZAKHAROV AND A. B. SHABAT, *A scheme for integrating the nonlinear equations of mathematical physics by the method of the inverse scattering problem*, Functional Anal. Appl., 8 (1974), pp. 393-404.

[16] X. ZHOU AND P. DEIFT, *Direct and inverse scattering transforms with arbitrary spectral singularities*, Comm. Pure Appl. Math., to appear.

[17] V. D. ZURKOWSKI, *Scattering for first order linear systems on the line and Bäcklund transformations*, Ph.D. Thesis, University of Minnesota, Minneapolis, MN, June, 1987.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 757-770, May 1990
© 1990 Society for Industrial and Applied Mathematics
011

ON THE SPREAD OF CONTINUOUS-TIME LINEAR SYSTEMS*

AVNER FRIEDMAN† AND MICHAEL L. HONIG‡

Abstract. Given the impulse response h of a linear time invariant system, this paper considers signals $y = h * u$ with inputs u subject to $|u(t)| \leqq 1$ and asks, for a given $\tau > 0$ and $y(t_0)$, what is the set of all the possible values (the "spread") of $y(t_0+\tau)$. This set is characterized, its properties are studied, and it is computed for some functions h.

Key words. control, input, output, signal, spread of linear systems

AMS(MOS) subject classifications. 93C60, 93B99, 94A99

1. Basic definitions and results. Let $h(t)$ be a prescribed continuous function defined for $0 \leqq t < \infty$ and belonging to $L^1(0, \infty)$; we refer to it as *an impulse response.* Let $u(t)$ be any measurable function for $0 \leqq t < \infty$ satisfying $|u(t)| \leqq 1$; we refer to it as an *input.* The function $y(t)$, defined

$$y(t) = \int_0^t h(t-s)u(s)\, ds,$$

is called the *output* or the *signal.*

Given $\alpha \in \mathbb{R}$, $t_0 \geqq 0$, $\tau > 0$, we would like to estimate the range of the output $y(t)$ at time $t = t_0 + \tau$, given that $y(t_0) = \alpha$. More quantitatively, we wish to bound the numbers

$$\tilde{\sigma}^+(\alpha, \tau, t_0) = \sup_u \{y(t_0+\tau);\ \text{given } y(t_0) = \alpha\},$$

$$\tilde{\sigma}^-(\alpha, \tau, t_0) = \inf_u \{y(t_0+\tau);\ \text{given } y(t_0) = \alpha\}.$$

Introduce the class of control functions

$$K_{\tau,\alpha} = \left\{ u \in L^\infty(-\infty, \tau),\ -1 \leqq u(s) \leqq 1,\ \int_{-\infty}^0 h(-s)u(s)\, ds = \alpha \right\} \tag{1.1}$$

and the functional

$$J_\tau(u) = \int_{-\infty}^\tau h(\tau - s)u(s)\, ds, \tag{1.2}$$

and define

$$\sigma^+(\tau, \alpha) = \sup_{u \in K_{\tau,\alpha}} J_\tau(u), \tag{1.3}$$

$$\sigma^-(\tau, \alpha) = \inf_{u \in K_{\tau,\alpha}} J_\tau(u). \tag{1.4}$$

* Received by the editors April 4, 1988; accepted for publication (in revised form) April 21, 1989. This work was partially supported by the National Science Foundation under grant DMS-86-12880.

† Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455.

‡ Bell Communications Research, Morristown, New Jersey 07960.

DEFINITION 1.1. The function $\sigma(\tau, \alpha) = \sigma^+(\tau, \alpha) - \sigma^-(\tau, \alpha)$ is called the *spread of the linear system.*

The motivation comes from the following theorem.

THEOREM 1.1. *For any* $\alpha \in \mathbb{R}$, $\tau > 0$,

$$\sup_{t_0 \geqq 0} \tilde{\sigma}^+(\tau, \alpha, t_0) = \sigma^+(\tau, \alpha), \tag{1.5}$$

$$\inf_{t_0 \geqq 0} \tilde{\sigma}^-(\tau, \alpha, t_0) = \sigma^-(\tau, \alpha), \tag{1.6}$$

and, consequently,

$$\sup_{t_0 \geqq 0} \tilde{\sigma}^+(\tau, \alpha, t_0) - \inf_{t_0 \geqq 0} \tilde{\sigma}^-(\tau, \alpha, t_0) = \sigma(\tau, \alpha). \tag{1.7}$$

Proof. The condition $y(t_0) = \alpha$ means that

$$\int_0^{t_0} h(t_0 - s)u(s)\, ds = \alpha. \tag{1.8}$$

Writing

$$y(t_0 + \tau) = \int_0^{t_0+\tau} h(t_0 + \tau - s)u(s)\, ds$$

and substituting $t_0 - s = -s'$, $u(t_0 + s') = v(s')$, we get

$$y(t_0 + \tau) = \int_{-t_0}^{\tau} h(\tau - s')v(s')\, ds'.$$

The same substitution applied to (1.8) gives

$$\int_{-t_0}^{0} h(-s')v(s')\, ds' = \alpha.$$

Hence

$$\tilde{\sigma}^+(\tau, \alpha, t_0) = \sup \left\{ \int_{-t_0}^{\tau} h(\tau - s')v(s')\, ds'; \quad v \text{ satisfies } |v(s')| \leqq 1, \right.$$
$$\left. \int_{-t_0}^{0} h(-s')v(s')\, ds' = \alpha \right\}.$$

Extending $v(s')$ to $s' < -t_0$ by zero, we see that $\sigma^+(\tau, \alpha, t_0)$ is $\sup J_\tau(v)$ when v is restricted to a subset say K_{τ,α,t_0}, of $K_{\tau,\alpha}$; hence

$$\tilde{\sigma}^+(\tau, \alpha, t_0) \leqq \sigma^+(\tau, \alpha).$$

As $t_0 \to \infty$ the subsets K_{τ,α,t_0} increase and every $u \in K_{\tau,\alpha}$ restricted to a bounded interval is a function in $\bigcup_{t_0>0} K_{\tau,\alpha,t_0}$ restricted to the same interval; this implies the equality in (1.5). The proof of (1.6) is similar.

THEOREM 1.2. *For any* $\alpha \in \mathbb{R}$, $\tau > 0$ *there exist admissible functions* $u^+_{\tau,\alpha}$, $u^-_{\tau,\alpha}$ *in* $K_{\tau,\alpha}$ *such that*

$$J_\tau(u^+_{\tau,\alpha}) = \sup_{u \in K_{\tau,\alpha}} J_\tau(u) = \sigma^+(\tau, \alpha), \tag{1.9}$$

$$J_\tau(u^-_{\tau,\alpha}) = \inf_{u \in K_{\tau,\alpha}} J_\tau(u) = \sigma^-(\tau, \alpha). \tag{1.10}$$

Indeed, taking a maximizing sequence u_j, we can extract a subsequence that is weakly convergent in L^1_{loc} to a function u_0. It is easy to check that u_0 is a maximizer for J_τ, i.e., u_0 is the asserted $u^+_{\tau,\alpha}$. The proof of (1.10) is similar.

In this paper we study the structure of $u^{\pm}_{\tau,\alpha}$ and this enables us to compute the spread of some linear systems of interest. In § 2 we solve a general maximization problem, which is then used in § 3 to analyze the structure of $u^{\pm}_{\tau,\alpha}$. In § 4 we establish various properties of $\sigma(\tau, \alpha)$, and in § 5 we compute $\sigma(\tau, \alpha)$ for some examples. Finally, in § 6 we show that all the results can be extended to the case where $y(t_0)$, $y(t_0+\tau_1), \cdots, y(t_0+\tau_{N-1})$ are prescribed and the range of $y(t_0+\tau_N)$ is sought; here $0<\tau_1<\tau_2<\cdots<\tau_N$.

Motivation for studying the function $\sigma^{\pm}(\tau, \alpha)$ comes from the following problem, posed in [1] and [2]. For any $d>0$, $T>0$ and impulse response $h(\cdot)$, denote by $N_{\max}(T, d)$ the maximum number of inputs $u_j(s)$ such that the corresponding outputs $y_j(t)$ satisfy

$$\max_{0<t\leqq T} |y_i(t)-y_j(t)| \geqq d \quad \forall i \neq j.$$

Since the mapping $u \to y$, in $L^\infty(0, T)$, maps the set of inputs u into a compact subset, the number $N_{\max}(T, d)$ is finite. We define

$$MCT(d) = \lim_{T\to\infty} \frac{\log N_{\max}(T, d)}{T} \text{ bits/sec}, \tag{1.11}$$

and would like to obtain bounds on the $MCT(d)$ for any $h(\cdot)$. Set

$$\tau^* = \inf\{\tau \mid \sigma(\tau, 0) = d\}.$$

Work in progress [3] indicates that

$$MCT(d) \leqq \frac{1}{\tau^*} \tag{1.12}$$

for any $h(\cdot)$ that satisfies

$$\int_{\{h(\tau-s)/h(-s)\geqq 1\}} |h(-s)|\, ds \leqq \int_{\{h(\tau-s)/h(-s)\leqq 1\}} |h(-s)|\, ds$$

for all $\tau \leqq \tau^*$; the arguments used depend on results derived in this paper. Results obtained here (in § 6) for the N constraint problem, in which N output values are specified, can be used to tighten the upper bound given by (1.12) (see [3]).

The problem of computing spread for a discrete-time linear system with impulse response h_i, $i=0, 1, 2, \cdots$, is considered in [2]. This computation is equivalent to solving a linear program with bounded variables and one equality constraint. Here we show how the spread can be computed for a continuous-time linear system. Two examples of special interest are presented in which the spread can be computed by finding a solution to a transcendental equation.

2. A general optimization problem. Let $f(s), g(s)$ be continuous functions in $\infty < s \leqq 0$ that belong to $L^1(-\infty, 0)$, and assume that

$$f \neq 0 \quad \text{a.e.}, \tag{2.1}$$

$$\text{meas}\left\{\frac{g}{f} = \mu\right\} = 0 \quad \text{for any } \mu \in \mathbb{R}. \tag{2.2}$$

Let

$$K=\left\{u(s) \text{ measurable for } -\infty<s<0, |u(s)|\leqq 1, \int_{-\infty}^{0} f(s)u(s)\,ds=\alpha\right\}$$

for some fixed $\alpha\in\mathbb{R}$, and

$$J(u)=\int_{-\infty}^{0} g(s)u(s)\,ds.$$

As in the proof of Theorem 1.2, we can show that there exists a function $u_0\in K$ such that

$$J(u_0)=\max_{u\in K} J(u). \tag{2.3}$$

THEOREM 2.1. *For any solution $u_0\in K$ of (2.3) there exists a number $\lambda\in\mathbb{R}$ such that almost everywhere*

$$u_0(s)=\begin{cases}\operatorname{sgn} f(s) & \text{if } g(s)/f(s)>\lambda,\\ -\operatorname{sgn} f(s) & \text{if } g(s)/f(s)<\lambda.\end{cases} \tag{2.4}$$

Note that (2.4) is equivalent to

$$u_0(s)=\operatorname{sgn}[g(s)-\lambda f(s)].$$

Proof. We begin by proving that $u_0=1$ almost everywhere. If the assertion is not true then the set $G_0=\{|u_0|<1\}$ has positive measure. Denote by G the subset of G_0 consisting of all points t of G_0-density equal to 1, such that also $f(t)\neq 0$. Then meas $G=$ meas $G_0>0$.

Take t_1, t_2 in $G(t_1\neq t_2)$ and let G_i be a subset of G contained in the δ_0-neighborhood of t_i, such that $\sup_{G_i}|u|<1$, meas $G_i\neq 0$ and $2\delta_0<|t_1-t_2|$. By decreasing one of these sets we arrive at the situation where

$$G_1\cap G_2=\varnothing, \qquad \text{meas } G_1=\text{meas } G_2=\delta>0.$$

For any real numbers A_1, A_2, if ε is positive and small enough then the function

$$\tilde{u}=u_0+A_1\frac{\varepsilon}{\delta}\chi_{G_1}+A_2\frac{\varepsilon}{\delta}\chi_{G_2} \tag{2.5}$$

satisfies $|\tilde{u}|\leqq 1$. Furthermore, if

$$A_1\int_{G_1} f(s)\,ds+A_2\int_{G_2} f(s)\,ds=0, \tag{2.6}$$

then $\int_{-\infty}^{0} f(s)\tilde{u}(s)\,ds=\alpha$, so that $\tilde{u}\in K$. Note that (2.6) is equivalent to

$$A_1 f(t_1)+A_2 f(t_2)=\sigma_1(\delta_0) \tag{2.7}$$

for some $\sigma_1(\delta_0)$ such that $\sigma_1(\delta_0)\to 0$ if $\delta_0\to 0$.

From the maximality of u_0 it follows that (2.6), or (2.7), implies $J(\tilde{u})\leqq J(u_0)$, that is,

$$A_1\int_{G_1} g(s)\,ds+A_2\int_{G_2} g(s)\,ds\leqq 0, \tag{2.8}$$

i.e.,

$$A_1 g(t_1)+A_2 g(t_2)\leqq \sigma_2(\delta_0) \tag{2.9}$$

for some $\sigma_2(\delta_0)$ such that $\sigma_2(\delta_0)\to 0$ if $\delta_0\to 0$.

If we choose

$$A_1 = -A_2 \frac{f(t_2)}{f(t_1)} + \frac{\sigma_1(\delta_0)}{f(t_1)} \tag{2.10}$$

so that (2.7) is satisfied, (2.9) must then hold and, upon letting $\delta_0 \to 0$, we get

$$A_2\left[-\frac{g(t_1)f(t_2)}{f(t_1)} + g(t_2)\right] \leqq 0. \tag{2.11}$$

Since A_2 is arbitrary, it follows that the expression in brackets must vanish. Thus

$$\frac{g(t_1)}{f(t_1)} = \frac{g(t_2)}{f(t_2)}$$

for all t_1, t_2 in G. Since G has a positive measure, this is a contradiction to (2.2).

Denote by D the set of all points t such that $f(t) \neq 0$ and t is a Lebesgue point of u_0. Thus almost all t in $(-\infty, 0)$ belong to D. Take any t_1, t_2 in D with

$$\frac{g(t_1)}{f(t_1)} > \frac{g(t_2)}{f(t_2)}. \tag{2.12}$$

We will prove that almost everywhere

$$u_0(t_2) = \operatorname{sgn} f(t_2) \text{ implies } u_0(t_1) = \operatorname{sgn} f(t_1), \tag{2.13}$$

$$u_0(t_1) = -\operatorname{sgn} f(t_1) \text{ implies } u_0(t_2) = -\operatorname{sgn} f(t_2). \tag{2.14}$$

These two statements clearly imply assertion (2.4).

To prove (2.13) suppose the assertion is not true. Then the set $\tilde{G}$ of the pair (t_1, t_2) for which (2.13) is not true has positive measure. Choose t_1, t_2 at which $\tilde{G}$ has density 1. Since t_1 and t_2 are Lebesgue points of the function $u_0(t)$ and $|u_0| = 1$ almost everywhere, for any $\delta_0 > 0$ we can find sets G_1, G_2 such that meas $G_i \neq 0$, G_i is contained in the δ_0-neighborhood of t_i, and

$$u_0(t) = \operatorname{sgn} f(t) \quad \text{for all } t \in G_2,$$
$$u_0(t) = -\operatorname{sgn} f(t) \quad \text{for all } t \in G_1.$$

By choosing $2\delta_0 < |t_1 - t_2|$ and by suitably decreasing one of the sets G_i, we get $G_1 \cap G_2 = \phi$, meas $G_1 =$ meas G_2. We again form the function (2.5). If

$$A_2 \operatorname{sgn} f(t_2) < 0, \qquad A_1 \operatorname{sgn} f(t_1) > 0, \tag{2.15}$$

then $|\tilde{u}| \leqq 1$ if ε is sufficiently small.

If we can further choose A_1, A_2 such that (2.6) (or (2.7)) holds, then (2.8) (or (2.9)) must be satisfied. Condition (2.7) is satisfied by the choice (2.10) of A_1, and if $A_2 \operatorname{sgn} f(t_2) < 0$, then clearly also $A_1 \operatorname{sgn} f(t_1) > 0$ provided δ_0 is sufficiently small. We conclude, after letting $\delta_0 \to 0$, that (2.11) must hold provided $A_2 \operatorname{sgn} f(t_2) < 0$. Dividing (2.11) by $A_2 f(t_2)$, we arrive at the inequality

$$-\frac{g(t_1)}{f(t_1)} + \frac{g(t_2)}{f(t_2)} \geqq 0,$$

which is a contradiction to (2.12). This completes the proof of (2.13); the proof of (2.14) is similar.

From Theorem 2.1 we immediately get Corollary 2.2.

COROLLARY 2.2. *The constant λ in Theorem 2.1 is uniquely determined by*

$$\int_{\{g(s)/f(s) > \lambda\}} |f(s)|\, ds - \int_{\{g(s)/f(s) < \lambda\}} |f(s)|\, ds = \alpha; \tag{2.16}$$

consequently the maximizer u_0 is also uniquely determined. As α decreases from $\int_{-\infty}^{0}|f(s)|\,ds$ to $-\int_{-\infty}^{0}|f(s)|\,ds$, $\lambda=\lambda(\alpha)$ increases monotonically from

$$\inf_{s<0}\{g(s)/f(s)\} \quad \text{to} \quad \sup_{s<0}\{g(s)/f(s)\}.$$

3. The structure of $u_{\tau,\alpha}^{\pm}$. *Choose $h(t)$* as in § 1, i.e.,

$$h\in L^1(0,\infty)\cap C^0[0,\infty) \tag{3.1}$$

and assume further that

$$h(t)\neq 0 \quad \text{a.e.}, \tag{3.2}$$

$$\operatorname{meas}\left\{0<t<\infty;\ \frac{h(t+\tau)}{h(t)}=\lambda\right\}=0 \quad \forall\tau>0, \quad \lambda\in\mathbb{R}. \tag{3.3}$$

Taking $f(t)=h(-t)$, $g(t)=h(\tau-t)$ in Theorem 2.1 and Corollary 2.2, we get Theorem 3.1.

THEOREM 3.1. *There exists a unique solution $u_{\tau,\alpha}^{+}$ of* (1.9) *given by*

$$u_{\tau,\alpha}^{+}(s)=\begin{cases}\operatorname{sgn} h(-s) & \text{if } \dfrac{h(\tau-s)}{h(-s)}>\lambda^{+},\\[2ex] -\operatorname{sgn} h(-s) & \text{if } \dfrac{h(\tau-s)}{h(-s)}<\lambda^{+}\end{cases} \tag{3.4}$$

where λ^{+} is determined by

$$\int_{\{h(\tau-s)/h(-s)>\lambda^{+}\}}|h(-s)|\,ds-\int_{\{h(t-s)/h(-s)<\lambda^{+}\}}|h(-s)|\,ds=\alpha. \tag{3.5}$$

Clearly also $u_{\tau,\alpha}^{+}(s)=\operatorname{sgn} h(\tau-s)$ if $0<s<\tau$.

We now consider a special case.

THEOREM 3.2. *If $h\in L^1(0,\infty)$, $h>0$, $d^2(\log h)/dt^2>0$, then there is a unique solution of* (1.9) *given by*

$$u_{\tau,\alpha}^{+}(s)=\begin{cases}1 & \text{if } -\infty<s<\mu,\\ -1 & \text{if } \mu<s<0\end{cases} \tag{3.6}$$

and $u_{\tau,\alpha}^{+}(s)=1$ if $0<s<\tau$, where μ is determined by

$$\int_{-\infty}^{\mu}h(-s)\,ds-\int_{\mu}^{0}h(-s)\,ds=\alpha. \tag{3.7}$$

Proof. By assumption,

$$\frac{h'(s)}{h(s)} \text{ is strictly increasing;}$$

hence

$$\frac{h'(\tau+s)}{h(\tau+s)}>\frac{h'(s)}{h(s)}.$$

This means that

$$\frac{d}{ds}\,\frac{h(\tau+s)}{h(s)}>0,$$

and thus

$$\frac{h(\tau-s)}{h(-s)} \text{ is strictly decreasing in } s.$$

Now apply Theorem 3.1 to complete the proof.

Remark 3.1. If $\log h$ is convex (but not satisfying $d^2(\log h)/dt^2>0$), then we can approximate it by a smooth function h_n with $d^2(\log h_n)/dt^2>0$. Applying Theorem 3.2 to the corresponding maximizers $u_{\tau,\alpha}^{h_n}$, we deduce that there is a maximizer $u_{\tau,\alpha}$ (for h) having the form (3.6), (3.7). There may be other maximizers; for instance, if $h(t)=e^{-t}$ then every $u\in K_{\tau,\alpha}$ is a maximizer. (Note that (3.3) does not hold for $h(t)=e^{-t}$.)

THEOREM 3.3. *If $h\in L^1(0,\infty)$, $h>0$, and $d^2(\log h)/dt^2<0$, then there is a unique solution of* (1.9) *given by*

$$u_{\tau,\alpha}^{+}(s)=\begin{cases}-1 & \text{if } -\infty<s<\tilde{\mu},\\ 1 & \text{if } \tilde{\mu}<s<0\end{cases} \tag{3.8}$$

and $u_{\tau,\alpha}^{+}(s)=1$ if $0<s<\tau$, where $\tilde{\mu}$ is determined by

$$-\int_{-\infty}^{\tilde{\mu}} h(-s)\,ds+\int_{\tilde{\mu}}^{0} h(-s)\,ds=\alpha. \tag{3.9}$$

Note that $d\mu/d\alpha>0$, $d\tilde{\mu}/d\alpha<0$, where $\mu=\mu(\alpha)$ and $\tilde{\mu}=\tilde{\mu}(\alpha)$ are defined by (3.7) and (3.9), respectively.

4. Properties of the spread. Theorem 3.1 implies that

$$\begin{aligned}\sigma^{+}(\tau,\alpha)&=\int_{\{h(\tau-s)/h(-s)>\lambda^{+}\}}[\operatorname{sgn} h(-s)]h(\tau-s)\,ds\\&\quad-\int_{\{h(\tau-s)/h(-s)<\lambda^{+}\}}[\operatorname{sgn} h(-s)]h(\tau-s)\,ds+\int_0^{\tau}|h(\tau-s)|\,ds\\&=\int_{\{h(\tau-s)/h(-s)>\lambda^{+}\}}\frac{h(\tau-s)}{h(-s)}|h(-s)|\,ds\\&\quad-\int_{\{h(\tau-s)/h(-s)<\lambda^{+}\}}\frac{h(\tau-s)}{h(-s)}|h(-s)|\,ds+\int_0^{\tau}|h(\tau-s)|\,ds\end{aligned} \tag{4.1}$$

where λ^{+} is determined by (3.5). Similarly, we can show that

$$\begin{aligned}\sigma^{-}(\tau,\alpha)&=-\int_{\{h(\tau-s)/h(-s)>\lambda^{-}\}}\frac{h(\tau-s)}{h(-s)}|h(-s)|\,ds\\&\quad+\int_{\{h(\tau-s)/h(-s)<\lambda^{-}\}}|h(-s)|\,ds-\int_0^{\tau}|h(\tau-s)|\,ds\end{aligned} \tag{4.2}$$

where λ^{-} is determined by

$$-\int_{\{h(\tau-s)/h(-s)>\lambda^{-}\}}|h(-s)|\,ds+\int_{\{h(\tau-s)/h(-s)<\lambda^{-}\}}|h(-s)|\,ds=\alpha. \tag{4.3}$$

As α decreases from $\int_0^{\infty}|h(s)|\,ds$ to $-\int_0^{\infty}|h(s)|\,ds$, $\lambda^{-}(\alpha)$ decreases monotonically from $\sup_{s<0}\{h(\tau-s)/h(-s)\}$ to $\inf_{s<0}\{h(t-s)/h(-s)\}$. Also, $\lambda^{-}(0)=\lambda^{+}(0)$.

Combining (4.1) and (4.2) gives the spread

$$\frac{1}{2}\sigma(\tau,\alpha)=\frac{1}{2}[\sigma^+(\tau,\alpha)-\sigma^-(\tau,\alpha)]$$

$$=\int_{\{h(\tau-s)/h(-s)>\lambda_M\}}\frac{h(\tau-s)}{h(-s)}|h(-s)|\,ds \tag{4.4}$$

$$-\int_{\{h(\tau-s)/h(-s)<\lambda_m\}}\frac{h(\tau-s)}{h(-s)}|h(-s)|\,ds+\int_0^\tau|h(\tau-s)|\,ds$$

where $\lambda_M=\max(\lambda^-,\lambda^+)$ and $\lambda_m=\min(\lambda^-,\lambda^+)$.

THEOREM 4.1. *There holds*

$$\frac{\partial\sigma^\pm(\tau,\alpha)}{\partial\alpha}=\lambda^\pm. \tag{4.5}$$

Proof. Since $\lambda^+(\alpha)$ is a monotonically decreasing function of α, we can write

$$\int_{\{h(\tau-s)/h(-s)>\lambda^++\Delta\lambda\}}|h(-s)|\,ds-\int_{\{h(\tau-s)/h(-s)<\lambda^++\Delta\lambda\}}|h(-s)|\,ds=\alpha-\Delta\alpha \tag{4.6}$$

where $\Delta\lambda$, $\Delta\alpha$ are positive. Subtracting (4.6) from (4.1) gives

$$2\int_{\{\lambda^+<h(\tau-s)/h(-s)<\lambda^++\Delta\lambda\}}|h(-s)|\,ds=\Delta\alpha. \tag{4.7}$$

From (4.1), (4.6), and (4.7),

$$\sigma^+(\tau,\alpha)-\sigma^+(\tau,\alpha-\Delta\alpha)=2\int_{\{\lambda^+<h(\tau-s)/h(-s)<\lambda^++\Delta\lambda\}}\frac{h(\tau-s)}{h(-s)}|h(-s)|\,ds$$

$$=[\lambda^++\varepsilon(\Delta\lambda)]\Delta\alpha$$

where $\varepsilon(\Delta\lambda)\to0$ as $\Delta\lambda\to0$. Letting $\Delta\alpha\to0$ gives $\partial\sigma^+/\partial\alpha=\lambda^+$. A similar argument shows that $\partial\sigma^-/\partial\alpha=\lambda^-$.

THEOREM 4.2. (i) *$\sigma^+(\tau,\alpha)$ is concave in α, $\sigma^-(\tau,\alpha)$ is convex in α, and thus $\sigma(\tau,\alpha)$ is concave in α:*

(ii) *$\sigma^\pm(\tau,\alpha)=-\sigma^\pm(\tau,-\alpha)$ and therefore $\sigma(\tau,\alpha)=\sigma(\tau,-\alpha)$,*

(iii) *$\partial\sigma(\tau,\alpha)/\partial\alpha\leqq0$ if $\alpha>0$.*

Proof. Assertion (i) follows immediately from Theorem 4.1 and the fact that $\partial\lambda^+/\partial\alpha$ ($\partial\lambda^-/\partial\alpha$) is negative (positive) for all α. Assertion (ii) is obvious from the definition of $\sigma^\pm$. Finally, since $\sigma(\tau,\alpha)$ is concave in α (by (i)) and $\partial\sigma(\tau,\alpha)/\partial\alpha=0$ at $\alpha=0$ (by (ii)), (iii) follows.

We now specialize to the case where either $\log h$ is convex, so that

$$\sigma^+(\tau,\alpha)=\int_{-\infty}^{\mu}h(\tau-s)\,ds-\int_{\mu}^{0}h(\tau-s)\,ds+\int_0^\tau h(s')\,ds' \tag{4.8}$$

where μ is determined by (3.7), or $\log h$ is concave so that

$$\sigma^+(\tau,\alpha)=-\int_{-\infty}^{\tilde\mu}h(\tau-s)\,ds+\int_{\tilde\mu}^{0}h(\tau-s)\,ds+\int_0^\tau h(s')\,ds' \tag{4.9}$$

where $\tilde\mu$ is determined by (3.9).

THEOREM 4.3. *If $h'<0$ and $\log h$ is convex or concave, then*

$$\frac{\partial\sigma^\pm(\tau,\alpha)}{\partial\tau}>0. \tag{4.10}$$

Proof. If $\log h$ is convex, then from (4.8) we get

$$\frac{\partial\sigma^+(\tau,\alpha)}{\partial\tau} = -\int_{-\infty}^{\mu}\frac{d}{ds}h(\tau-s)\,ds+\int_{\mu}^{0}\frac{d}{ds}h(\tau-s)\,ds+h(\tau)$$

$$=2h(\tau)-2h(\tau-\mu)>0.$$

Similarly, if $\log h$ is concave then

$$\frac{\partial\sigma^+(\tau,\alpha)}{\partial\tau} = \int_{-\infty}^{\tilde{\mu}}\frac{d}{ds}h(\tau-s)\,ds-\int_{\tilde{\mu}}^{0}\frac{d}{ds}h(\tau-s)\,ds+h(\tau)$$

$$=2h(\tau-\tilde{\mu})>0.$$

Finally, the second inequality in (4.10) follows from the first inequality and Theorem 4.2(ii).

5. Examples. If $h(t)=\exp\{-k(t)\}$, where $k(t)\to\infty$, k convex (k concave), then $\log h$ is concave (convex). For $h(t)=(t+a)^b$ where $a>0$, $b>0$, $\log h$ is convex.

We now consider two functions $h(t)$ of special interest.

THEOREM 5.1. *Let*

$$h(t)=\sum_{i=1}^{N}a_i e^{-\beta_i t} \qquad (a_i>0,\ \beta_i>0). \tag{5.1}$$

Then $d^2\log h/dt^2>0$.

Proof. As in the proof of Theorem 3.2, the assertion is equivalent to showing that

$$\frac{d}{ds}\frac{h(\tau-s)}{h(-s)}=\frac{h(-s)\Sigma a_i\beta_i e^{-\beta_i(\tau-s)}-h(\tau-s)\Sigma a_i\beta_i e^{\beta_i s}}{h^2(-s)}$$

is negative for any $\tau>0$. But the numerator is equal to

$$\sum\sum a_i\beta_i a_j(e^{-\beta_i(\tau-s)+\beta_j s}-e^{-\beta_j(\tau-s)+\beta_i s})$$

$$=\sum\sum a_i a_j\beta_i e^{s(\beta_i+\beta_j)}(e^{-\beta_i\tau}-e^{-\beta_j\tau})$$

$$=\tfrac{1}{2}\sum\sum a_i a_j e^{s(\beta_i+\beta_j)}[\beta_i(e^{-\beta_i\tau}-e^{-\beta_j\tau})+\beta_j(e^{-\beta_j\tau}-e^{-\beta_i\tau})]$$

$$=\tfrac{1}{2}\sum\sum a_i a_j e^{s(\beta_i+\beta_j)}(\beta_i-\beta_j)(e^{-\beta_i\tau}-e^{-\beta_j\tau})$$

and each term in the last sum is negative if $\beta_i\neq\beta_j$.

For the function (5.1), the μ determined by (3.7) is given by

$$\sum_{i=1}^{N}\frac{a_i}{\beta_i}(2e^{\beta_i\mu}-1)=\alpha.$$

The next example is

$$h(t)=e^{-\beta t}\cos\omega t \qquad (\beta>0,\ \omega>0). \tag{5.2}$$

Since

$$\frac{h(\tau-s)}{h(-s)}=e^{-\beta\tau}(\cos\omega\tau+\sin\omega\tau\tan\omega s),$$

we can check that the optimal solution $u^+_{\tau,\alpha}$, which for simplicity we will denote by u_0, satisfies

$$u_0(s)=\begin{cases}\operatorname{sgn} h(-s) & \text{if } \gamma-n\pi<\omega s<-\dfrac{(2n+1)\pi}{2},\\[2ex] -\operatorname{sgn} h(-s) & \text{if } -\dfrac{(2n+3)\pi}{2}<\omega s<\gamma-n\pi\end{cases}$$

if $n=0,1,2,\cdots$, and

$$u_0(s)=\begin{cases}-\operatorname{sgn} h(-s) & \text{if } -\pi/2<\omega s<\min(\gamma+\pi,0),\\ \operatorname{sgn} h(-s) & \text{if } \min(\gamma+\pi,0)<\omega s\leqq 0\end{cases}$$

where $\gamma\in[-3\pi/2,-\pi/2]$ is to be selected such that

$$\int_{-\infty}^{0} h(-s)u_0(s)\,ds=\alpha. \tag{5.3}$$

Recalling (5.2) we can check that

$$u_0(s)=\begin{cases}-1 & \text{if } \gamma-2n\pi<\omega s<\gamma-(2n-1)\pi,\\ 1 & \text{if } \gamma-(2n-1)\pi<\omega s<\gamma-2(n-1)\pi\end{cases}$$

for $n=1,2,\cdots$, and

$$u_0(s)=\begin{cases}-1 & \text{if } \gamma<\omega s<\min(\gamma+\pi,0),\\ 1 & \text{if } \min(\gamma+\pi,0)<\omega s<0.\end{cases}$$

Setting $\gamma'=\min(\gamma+\pi,0)$ and using the formula

$$\int_a^b h(-s)\,ds=-\operatorname{Re}\left\{\frac{1}{\beta+i\omega}\left[e^{-(\beta+i\omega)b}-e^{-(\beta+i\omega)a}\right]\right\},$$

we can compute

$$\int_{-\infty}^{0} h(-s)u_0(s)\,ds=\sum_{n=1}^{\infty}\left[-\int_{(\gamma-2n\pi)/\omega}^{(\gamma-(2n-1)\pi)/\omega} h(-s)\,ds+\int_{(\gamma-(2n-1)\pi)/\omega}^{(\gamma-2(n-1)\pi)/\omega} h(-s)\,ds\right]$$
$$-\int_{\gamma/\omega}^{\gamma'/\omega} h(-s)\,ds+\int_{\gamma'/\omega}^{0} h(-s)\,ds.$$

After somewhat lengthy calculations we get the expression

$$\operatorname{Re}\left\{\frac{\beta-i\omega}{\beta^2+\omega^2}\,e^{(\beta+i\omega)\gamma/\omega}\,\frac{1+e^{-\beta\pi/\omega}}{1-e^{-\beta\pi/\omega}}+\frac{\beta-i\omega}{\beta^2+\omega^2}\left[1-2\,e^{(\beta+i\omega)\gamma'/\omega}+e^{(\beta+i\omega)\gamma/\omega}\right]\right\}, \tag{5.4}$$

or

$$\frac{\beta}{\beta^2+\omega^2}+\frac{2\,e^{\beta\gamma/\omega}}{(\beta^2+\omega^2)(1-e^{-\beta\pi/\omega})}(\beta\cos\gamma+\omega\sin\gamma)-\frac{2\,e^{\beta\gamma'/\omega}}{\beta^2+\omega^2}(\beta\cos\gamma'+\omega\sin\gamma').$$

Hence (5.3) determines γ by the following formulas:

$$\frac{2\,e^{\beta\gamma/\omega}}{1-e^{-\beta\pi/\omega}}(\beta\cos\gamma+\omega\sin\gamma)=\alpha(\omega^2+\beta^2)+\beta \quad \text{if } -\pi<\gamma<0,$$

$$\left[\frac{2\,e^{\beta\gamma/\omega}}{1-e^{-\beta\pi/\omega}}+2\,e^{\beta(\gamma+\pi)/\omega}\right](\beta\cos\gamma+\omega\sin\gamma) \tag{5.5}$$

$$=\alpha(\omega^2+\beta^2)-\beta \quad \text{if } -3\pi/2<\gamma<-\pi.$$

Since

$$\int_{-\infty}^{0} h(\tau-s)u_0(s)\,ds=\operatorname{Re}\left\{\int_{-\infty}^{0} e^{-(\beta+i\omega)(\tau-s)}u_0(s)\,ds\right\}$$

$$=\operatorname{Re}\left\{e^{-(\beta+i\omega)\tau}\int_{-\infty}^{0} h_0(-s)u_0(s)\,ds\right\}$$

and the last integral is equal to the expression in braces in (5.4), we find that

$$\sigma^+(\tau,\alpha)=\frac{2\,e^{\beta((\gamma/\omega)-\tau)}}{(\beta^2+\omega^2)(1-e^{-\beta\pi/\omega})}[\beta\cos(\gamma-\omega\tau)+\omega\sin(\gamma-\omega\tau)]$$

$$-\frac{e^{-\beta\tau}}{\beta^2+\omega^2}(\beta\cos\omega\tau-\omega\sin\omega\tau)+\int_0^{\tau}|h(s)|\,ds \quad \text{if } -\pi<\gamma<0,$$

$$\sigma^+(\tau,\alpha)=\frac{2\,e^{\beta((\gamma+\pi)/\omega-\tau)}}{(\beta^2+\omega^2)(1-e^{-\beta\pi/\omega})}[\beta\cos(\gamma-\omega\tau)+\omega\sin(\gamma-\omega\tau)]$$

$$+\frac{e^{-\beta\tau}}{\beta^2+\omega^2}(\beta\cos\omega\tau-\omega\sin\omega\tau)+\int_0^{\tau}|h(s)|\,ds \quad \text{if } -\frac{3\pi}{2}<\gamma<-\pi.$$

6. Several constraints. The results of the previous sections can be extended to the case of several constraints. In fact it all hinges on generalizing Theorem 2.1 to the problem

$$\max_{u\in K_\alpha}\int_{-\infty}^{0} g(s)u(s)\,ds \tag{6.1}$$

where K_α is the set of all measurable functions $u(s)$ satisfying

$$-1\leqq u(s)\leqq 1 \quad \text{for } -\infty<s\leqq 0, \tag{6.2}$$

$$\int_{-\infty}^{0} f_i(s)u(s)\,ds=\alpha_i \qquad (i=1,2,\cdots,N). \tag{6.3}$$

Here g and f_i are given functions in $L^1(-\infty,0)\cap C^0(-\infty,0]$ and α_i are given real numbers.

THEOREM 6.1. *Assume that $f_1\neq 0$ almost everywhere and that, for any real numbers $\mu_1,\cdots,\mu_N$,*

$$\text{measure}\left\{g=\sum_{i=1}^{N}\mu_i f_i\right\}=0.$$

Then there exist sequences u_m, $\lambda_{i,m}$, $\alpha_{i,m}$ with $u_m\to u_0$ weakly in L^1_{loc}, $\alpha_{1,m}=\alpha_1$, $\alpha_{i,m}\to\alpha_i$ for $2\leqq i\leqq N$, where u_0 is a maximizer of (6.1), *and*

$$u_m(s)=\operatorname{sgn}\left[g(s)-\sum_{i=1}^{N}\lambda_{i,m}f_i(s)\right], \tag{6.4}$$

$$\int_{-\infty}^{0} f_i(s)u_m(s)\,ds = \alpha_{i,m} \qquad (i=1,2,\cdots,N). \tag{6.5}$$

Thus to evaluate (6.1) we need to analyze the u_m from (6.4), (6.5) and then compute $\int_{-\infty}^{0} gu_m$, noting that

$$\int_{-\infty}^{0} g(s)u_m(s)\,ds \to \int_{-\infty}^{0} g(s)u_0(s)\,ds = \max_{u\in K_\alpha} \int_{-\infty}^{0} g(s)u(s)\,ds.$$

Proof. For any small $\eta > 0$ introduce the "penalized" functional

$$J_\eta(u) = \int_{-\infty}^{0} g(s)u(s)\,ds - \frac{1}{\eta}\sum_{i=2}^{N}\left[\int_{-\infty}^{0} f_i(s)u(s)\,ds - \alpha_i\right]^2 \tag{6.6}$$

and consider the problem

$$\text{maximize } J_\eta(u) \quad \text{for } u\in K \tag{6.7}$$

where K consists of all functions u satisfying

$$-1 \leqq u(s) \leqq 1, \qquad \int_{-\infty}^{0} f_1(s)\,ds = \alpha_1.$$

Proceeding as in the proof of Theorem 2.1, we deduce that if $|\tilde{u}| \leqq 1$, where $\tilde{u}$ is defined by (2.5) with $u_0 = u_\eta$, then (2.6) implies

$$A_1 g(t_1) + A_2 g(t_2) - \frac{2}{\eta}\sum_{i=2}^{N}\left(\int_{-\infty}^{0} f_i u_\eta\,ds - \alpha_i\right)(A_1 f_i(t_1) + A_2 f_i(t_2)) \leqq \sigma_2(\delta_0)$$

where u_η is a solution of (6.7) and $\sigma_2(\delta_0)\to 0$ if $\delta_0 \to 0$. Taking $\delta_0 \to 0$, we get the inequality

$$A_1\left[g(t_1) - \sum_{i=2}^{N}\lambda_{i,\eta} f_i(t_1)\right] + A_2\left[g(t_2) - \sum_{i=2}^{N}\lambda_{i,\eta} f_i(t_2)\right] \leqq 0$$

for some scalars $\lambda_{i,\eta}$. We can now proceed as in § 2 to deduce that meas $\{|u_\eta| < 1\} = 0$; furthermore,

$$u_\eta(s) = \operatorname{sgn}\left[g(s) - \sum_{i=1}^{N}\lambda_{i,\eta} f_i(s)\right]. \tag{6.8}$$

We note that

$$J_\eta(u_\eta) \geqq J_\eta(\hat{u}) \quad \forall\, \hat{u}\in K;$$

from this inequality it follows that

$$\frac{1}{\eta}\sum_{i=2}^{N}\left[\int_{-\infty}^{0} f_i(s)u_\eta(s)\,ds - \alpha_i\right]^2 \leqq C, \quad C \text{ independent of } \eta.$$

Hence, as $\eta \to 0$,

$$\alpha_{i,\eta} \equiv \int_{-\infty}^{0} f_i(s)u_\eta(s)\,ds \to \alpha_i \qquad (2 \leqq i \leqq N). \tag{6.9}$$

It is also easy to verify that for any convergent subsequence u_{η_m} (weakly in L^1_{loc}), the limit u_0 is a solution to problem (6.1). Indeed

$$J_\eta(u_\eta) \leqq \int_{-\infty}^{0} gu_\eta\,ds \leqq \max_{u\in K_{\alpha_\eta}} \int_{-\infty}^{0} gu\,ds = \int_{-\infty}^{0} g\hat{u}_\eta\,ds \tag{6.10}$$

where K_{α_η} is defined as K_α but with

$$\alpha_2 = \alpha_{2,\eta}, \cdots, \alpha_N = \alpha_{N,\eta}.$$

Since $\alpha_{j,\eta} \to \alpha_j$, if we take η to vary in a subsequence of η_m such that $\hat{u}_\eta \to \hat{u}$ weakly in L^1_{loc}, then $\int_{-\infty}^0 g\hat{u}_\eta \to \int_{-\infty}^0 g\hat{u}$ and $\hat{u} \in K_\alpha$ (i.e., $\hat{u}$ satisfies (6.2), (6.3)). Denoting by u_1 any solution of (6.1)-(6.3), we then have

$$\int_{-\infty}^0 g\hat{u}\, ds \leqq \int_{-\infty}^0 gu_1\, ds;$$

also, by maximality of u_η (see (6.7)),

$$\int_{-\infty}^0 gu_1\, ds = J_\eta(u_1) \leqq J_\eta(u_\eta).$$

Using these relations in (6.10) and noting that

$$\int_{-\infty}^0 gu_\eta\, ds \to \int_{-\infty}^0 gu_0\, ds,$$

we conclude that

$$\int_{-\infty}^0 gu_0\, ds = \int_{-\infty}^0 gu_1\, ds = \max_{u \in K_\alpha} \int_{-\infty}^0 gu\, ds.$$

Thus u_0 is a solution to (6.1)-(6.3). Recalling (6.8), (6.9) completes the proof of Theorem 6.1.

Remark 6.1. The $\lambda_{i,m}$ satisfy

$$\int_{\{g > \Sigma\lambda_{j,m}f_j\}} f_i(s)\, ds - \int_{\{g < \Sigma\lambda_{j,m}f_j\}} f_i(s)\, ds = \alpha_i \qquad (1 \leqq i \leqq N).$$

From these equations we should be able to determine the $\lambda_{j,m}$, at least in some relatively simple examples, and show that $\lambda_{j,m} \to \lambda_j$ (λ_j finite) as $m \to \infty$; this would imply that

$$u_0 = \operatorname{sgn}\left[g(s) - \sum_{i=1}^N \lambda_i f_i(s)\right],$$

$$\int_{-\infty}^0 f_j(s)u_0(s)\, ds = \alpha_j \quad \text{for } 1 \leqq j \leqq N.$$

Remark 6.2. Theorem 2.1 can actually also be proved using the penalized functional

$$\int_{-\infty}^0 gu\, ds - \frac{1}{\eta}\left(\int_{-\infty}^0 fu\, ds - \alpha\right)^2 - \int_{-\infty}^0 \frac{(u-u_0)^2}{1+s^2}\, ds.$$

Remark 6.3. Consider the problem

$$\max_{u \in K} J_\tau(u) \tag{6.11}$$

where K is the set of all inputs u that satisfy

$$\int_{-\infty}^{t_j} h(t_j - s)u(s)\, ds = \alpha_j, \qquad j = 1, \cdots, N \tag{6.12}$$

and where $J_\tau(u)$ is defined by (1.2) and $0 = t_1 < t_2 < \cdots < t_{N-1} < t_N \equiv \tau$.

Set

$$\sigma^+(\tau; t_1, \alpha_1, \cdots, t_N, \alpha_N) = \max_{u \in K} J_\tau(u), \tag{6.13a}$$

$$\sigma^-(\tau; t_1, \alpha_1, \cdots, t_N, \alpha_N) = \min_{u \in K} J_\tau(u). \tag{6.13b}$$

Then there exists a solution to (6.11) if and only if

$$
\begin{aligned}
&|\alpha_1| \leqq \int_0^\infty |h(s)|\, ds, \\
&\sigma^-(t_j; t_1, \alpha_1, \cdots, t_{j-1}, \alpha_{j-1}) \leqq \alpha_j \leqq \sigma^+(t_j; t_1, \alpha_1, \cdots, t_{j-1}, \alpha_{j-1}), \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad 1 < j \leqq N.
\end{aligned} \tag{6.14}
$$

If we assume $h(s) = 0$ for $s < 0$, the conclusion of Remark 6.1 becomes

$$
u(s) = \begin{cases} \operatorname{sgn}\left[h(\tau - s) - \sum_{i=1}^{j} \lambda_i h(t_i - s) \right], & t_j < s < t_{j+1}, \\ \operatorname{sgn}\left[h(\tau - s) - \sum_{i=1}^{N} \lambda_i h(t_i - s) \right], & s < t_1 \end{cases} \tag{6.15}
$$

where $\lambda_1, \cdots, \lambda_N$ satisfy (6.12).

REFERENCES

[1] M. L. HONIG, S. BOYD, AND B. GOPINATH, *On optimum signal sets for digital communications with finite precision and amplitude constraints*, in Proc. IEEE Globecom. Conference, Tokyo, November 1987.

[2] M. L. HONIG, K. STEIGLITZ, AND B. GOPINATH, *Bounds on maximum throughput for digital communications with finite-precision and amplitude constraints*, in Proc. Internat. Conf. Acoustics, Speech, and Signal Processing, New York, NY, April 1988.

[3] M. L. HONIG, K. STEIGLITZ, S. BOYD, AND B. GOPINATH, *Bounds on maximum throughput for digital communications with finite-precision and amplitude constraints*, IEEE Trans. Inform. Theory, to appear.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 771-781, May 1990
© 1990 Society for Industrial and Applied Mathematics
012

THE EXPANSION OF A HOLOMORPHIC FUNCTION IN A LAPLACE SERIES*

HANS VOLKMER†

Abstract. It is shown that a holomorphic function defined on a suitable subset of the complex manifold $\{(x, y, z) \in \mathbb{C}^3: x^2+y^2+z^2=1\}$ can be expanded in a series of spherical surface harmonics. When sphero-conal coordinates are introduced, this result gives the expansion of a holomorphic function of two variables in a series of products of Lamé polynomials.

Key words. spherical harmonics, Lamé polynomials, Laplace series

AMS(MOS) subject classifications. 33A45, 33A55

Introduction. In 1862 K. Neumann [6] proved that every function holomorphic inside an ellipse with foci at the points ± 1 can be expanded in a locally uniformly convergent series of Legendre polynomials. Later Hobson [4, Chap. VII] and others studied the convergence of such a Legendre series associated with arbitrary functions defined on the focal line $[-1, 1]$. It is not surprising that the latter results from real analysis are more difficult to state and prove than the original theorem from complex analysis.

Now it is well known that a Legendre series is a particular case of a Laplace series, i.e., the expansion of a function defined on the two-dimensional unit sphere

$$S = \{(x, y, z) \in \mathbb{R}^3: x^2+y^2+z^2=1\}$$

in a series of spherical surface harmonics. The convergence properties of Laplace series are thoroughly studied in the framework of real analysis, see Hobson [4, Chap. VII] or Sansone [7, Chap. III]. However, there seems to be no result in the literature extending Neumann's expansion theorem to Laplace series. In this paper we will prove such a theorem on the expansion of a holomorphic function in a Laplace series.

To formulate our theorem we must first specify domains that replace the domains bounded by ellipses in Neumann's theorem. These domains are subsets of the two-dimensional complex manifold

$$T = \{(x, y, z) \in \mathbb{C}^3: x^2+y^2+z^2=1\}$$

and are defined by

$$T_\gamma = \{(x, y, z) \in T: |x|^2+|y|^2+|z|^2 < \cosh 2\gamma\}$$

where γ is any positive real number. The system of sets T_γ is nested in the sense

$$S \subset T_\beta \subset T_\gamma \subset T \quad \text{if } 0 < \beta < \gamma < \infty.$$

This situation is analogous to that of the nested system of interior domains of ellipses with foci at the points ± 1. In this analogy the unit sphere S corresponds to the focal line $[-1, 1]$. Our main Theorem 4.7 in § 4 then states that every function holomorphic on T_γ can be expanded in a locally uniformly convergent Laplace series. This Laplace series is the ordinary one associated with the given function restricted to S. Of course, the spherical surface harmonics involved in Laplace series have to be continued holomorphically from S to T in order to make our statement meaningful.

* Received by the editors September 6, 1988; accepted for publication (in revised form) July 25, 1989.
† Fachbereich Mathematik, Universität Gesamthochschule Essen, Universitätsstrasse 3, D 4300 Essen 1, Federal Republic of Germany.

The proof of the main theorem is simple. It uses Neumann's theorem and some well-known facts on the ordinary Laplace series. For convenience of the reader we have collected these basic results in §§ 1 and 2, respectively. In § 3 we derive some elementary properties of the manifold T and the sets T_γ which are needed to prove the main theorem in § 4.

If we introduce a complex version of sphero-conal coordinates in the manifold T, then our expansion theorem yields the expansion of a holomorphic function of two complex variables in a series of products of Lamé polynomials, see Theorem 5.2. An expansion of this kind has already been given by Volk [10] in 1925.

I am indebted to Prof. D. Schmidt who drew my attention to Volk's paper. Whereas Volk's results turned out to be incorrect they were the starting point for the work leading to the present paper.

1. Proof of Neumann's theorem. In this section we reprove Neumann's expansion theorem. The proof is adapted from Whittaker and Watson [12, Chap. XV]. We also refer to Szegö [9, Thm. 9.1.1] and Meixner and Schäfke [5, Thm. 3, p. 258] where Neumann's theorem appears as a special case of more general expansion theorems.

Let P_n be the Legendre polynomial of degree n. We know that the normalized polynomials $\sqrt{2n+1}\cdot P_n$, $n=0,1,2,\cdots$, form an orthonormal basis of the Hilbert space $L^2(I)$ of square-integrable functions on the interval $I=[-1,1]$ subject to the inner product

$$\langle f, g\rangle_I := \tfrac{1}{2}\int_{-1}^{1} f(t)\overline{g(t)}\,dt.$$

We mention that the completeness of the system $\sqrt{2n+1}\cdot P_n$ is a consequence of the fact that the space of polynomials is dense in $L^2(I)$. It follows that every function $f\in L^2(I)$ can be expanded in the L^2-convergent series

$$f(t)\sim \sum_{n=0}^{\infty}(2n+1)\langle f, P_n\rangle_I P_n(t). \tag{1.1}$$

This series is called the Legendre series associated with f. In order to prove Neumann's theorem, we now assume that f is a function holomorphic on the domain

$$E_\gamma := \{\cos\theta\colon \theta\in\mathbb{C}, |\operatorname{Im}\theta|<\gamma\},\qquad 0<\gamma\leqq\infty,$$

which is the interior of an ellipse with foci at the points ±1 and half-axes $\cosh\gamma$ and $\sinh\gamma$. If $\gamma=\infty$ then E_γ is the whole complex plane. We first derive a bound for the Legendre polynomial P_n using its Laplace's integral representation

$$P_n(\cos\theta)=\frac{1}{2\pi}\int_{-\pi}^{\pi}(\cos\theta+i\sin\theta\cos\varphi)^n\,d\varphi,\qquad \theta\in\mathbb{C}. \tag{1.2}$$

Since

$$|\cos\theta+it\sin\theta|\leqq\max(|e^{i\theta}|,|e^{-i\theta}|)=e^{|\operatorname{Im}\theta|}\quad\text{for } -1\leqq t\leqq 1,$$

(1.2) yields $|P_n(\cos\theta)|\leqq e^{n|\operatorname{Im}\theta|}$. This implies immediately

$$|P_n(z)|\leqq e^{n\alpha}\quad\text{for } z\in\bar{E}_\alpha,\qquad 0\leqq\alpha<\infty, \tag{1.3}$$

where $\bar{E}_\alpha$ is the closure of E_α and $\bar{E}_0:=I$.

Now consider the Legendre function Q_n of the second kind which is holomorphic on $\mathbb{C}\setminus I$ and satisfies

$$Q_n(t-i0)-Q_n(t+i0)=\pi i P_n(t),\qquad -1<t<1.$$

Since f is holomorphic on E_γ, this equation together with Cauchy's integral theorem shows that

$$\langle f, P_n\rangle_I = \frac{1}{2\pi i}\int_{\partial E_\beta} f(z)Q_n(z)\,dz, \qquad 0<\beta<\gamma, \tag{1.4}$$

where the integral is taken round the circumference ∂E_β of E_β in a positive direction. To estimate Q_n we use again its Laplace's integral representation

$$Q_n(\cosh w) = \int_0^\infty (\cosh w + \sinh w \cosh u)^{-n-1}\,du, \qquad \operatorname{Re} w > 0. \tag{1.5}$$

We write $w = w_1 + iw_2$, $w_1 > 0$, $w_2 \in \mathbb{R}$, and use the inequality

$$\begin{aligned}&|\cosh w + \sinh w \cosh u|^2\\ &\quad = (\cosh w_1 + \sinh w_1 \cosh u)^2 \cos^2 w_2 + (\sinh w_1 + \cosh w_1 \cosh u)^2 \sin^2 w_2\\ &\quad \geqq \max\,(e^{2w_1}, \sinh^2 w_1 \cosh^2 u).\end{aligned}$$

Then (1.5) gives

$$|Q_n(\cosh w)| \leqq e^{-nw_1}\frac{1}{\sinh w_1}\int_0^\infty \frac{du}{\cosh u}.$$

The above integral is equal to $\pi/2$. Hence we obtain from (1.4)

$$\begin{aligned}|\langle f, P_n\rangle_I| &\leqq \frac{|\partial E_\beta|}{2\pi} e^{-n\beta}\frac{1}{\sinh\beta}\frac{\pi}{2}\|f|\bar{E}_\beta\|_\infty\\ &\leqq \frac{\pi}{2}\coth\beta\, e^{-n\beta}\|f|\bar{E}_\beta\|_\infty.\end{aligned} \tag{1.6}$$

We used $\|f|\bar{E}_\beta\|_\infty$ to denote the maximum of $|f(z)|$ over $z \in \bar{E}_\beta$. Combining (1.3) and (1.6) we see that the Legendre series (1.1) has the convergent majorant

$$\frac{\pi}{2}\coth\beta\|f|\bar{E}_\beta\|_\infty \sum_{n=0}^\infty (2n+1)\,e^{n(\alpha-\beta)} \tag{1.7}$$

on $\bar{E}_\alpha$ whenever $0 \leqq \alpha < \beta < \gamma$. Hence the Legendre series associated with f converges normally on compact subsets of E_γ, i.e., it has a convergent majorant on compact subsets of E_γ. In particular, the Legendre series is locally uniformly convergent on E_γ. Hence, by Weierstrass's theorem, the sum of the series is holomorphic on E_γ. This sum is equal to f on I by the completeness of the system $\sqrt{2n+1}\cdot P_n$ in $L^2(I)$, consequently, it is equal to f on E_γ by the identity theorem. We have thus proved Neumann's theorem.

THEOREM 1.8. *Let f be a function holomorphic on E_γ where $0<\gamma\leqq\infty$. Then the Legendre series associated with f converges normally on compact subsets of E_γ to the sum f.*

In § 4 we will not use this theorem directly, but instead inequalities (1.3) and (1.6) which led to the majorant (1.7).

2. The Laplace series associated with functions on S. Concerning Laplace series we refer to Hobson [4, §§ 95 and 211], Sansone [7, Chap. III, §§ 18–24], and Schäfke [8, § 5.5]. We will need the following definitions and remarks.

A spherical harmonic of degree n is a polynomial in the variables x, y, z with complex coefficients which is homogeneous in x, y, z of degree n and which is harmonic, i.e., satisfies Laplace's equation. A spherical surface harmonic of degree n is a function defined on the two-dimensional unit sphere S, which is the restriction of a spherical harmonic of degree n onto S. The spherical surface harmonics of degree n form a complex linear space of dimension $2n+1$. Let $L^2(S)$ denote the Hilbert space of square-integrable functions on S endowed with the inner product

$$\langle f, g\rangle_S = \frac{1}{4\pi}\int_S f(p)\overline{g(p)}\,dp, \tag{2.1}$$

where the integral is taken over the sphere S of area 4π. Spherical surface harmonics of different degrees are orthogonal with respect to $\langle\ ,\ \rangle_S$. Moreover, spherical surface harmonics are complete in $L^2(S)$, i.e., every function $f \in L^2(S)$ can be expanded in the L^2-convergent series

$$f(p) \sim \sum_{n=0}^{\infty} f_n(p), \qquad p \in S, \tag{2.2}$$

where f_n is the orthogonal projection of f onto the linear space of spherical surface harmonics of degree n. The series (2.2) is called the Laplace series associated with f. The spherical surface harmonic f_n can be represented as an integral

$$\begin{aligned} f_n(p) &= \frac{2n+1}{4\pi}\int_S P_n(\langle p, p'\rangle) f(p')\,dp' \\ &= (2n+1)\langle f, P_n(\langle p, \cdot\rangle)\rangle_S, \end{aligned} \tag{2.3}$$

where P_n is the Legendre polynomial of degree n and $\langle\ ,\ \rangle$ denotes the usual inner product in $\mathbb{R}^3$.

Formula (2.3) will be used to derive estimates for f_n. First, we note that the norm of $P_n(\langle p, \cdot\rangle)$ with respect to the inner product $\langle\ ,\ \rangle_S$ is equal to $(2n+1)^{-1/2}$. This can be calculated by using polar coordinates in S with north pole p. Hence the Cauchy–Schwarz inequality applied to (2.3) gives

$$\|f_n\|_\infty \leqq \sqrt{2n+1}\,\|f\|_2 \tag{2.4}$$

where $\|f_n\|_\infty$ denotes the maximum of $|f_n(p)|$ over $p \in S$ and $\|f\|_2 = \langle f, f\rangle_S^{1/2}$. Obviously, (2.3) and thus (2.4) remain valid if f is replaced by f_n. It follows that

$$\|f_n\|_2 \leqq \|f_n\|_\infty \leqq \sqrt{2n+1}\,\|f_n\|_2. \tag{2.5}$$

This inequality means that the L^2-norm of a spherical surface harmonic cannot be "essentially" smaller than its max-norm.

Let us call the quantity

$$\rho(f) := \left(\limsup_{n\to\infty} \|f_n\|_\infty^{1/n}\right)^{-1} \tag{2.6}$$

the harmonic radius of f. We always have $\rho(f) \geqq 1$ by (2.4) and, by (2.5),

$$\rho(f) = \left(\limsup_{n\to\infty} \|f_n\|_2^{1/n}\right)^{-1}. \tag{2.7}$$

The harmonic radius of f will play the role of a convergence radius of the Laplace series associated with f in § 4. At this point we can give a potential theoretic interpretation of the harmonic radius which, however, will not be used in the sequel. Let (2.2) be the Laplace series associated with a function $f \in L^2(S)$. Then form the function

$$\hat{f}(rp) := \sum_{n=0}^{\infty} r^n f_n(p), \qquad 0 \leqq r < \rho(f), \quad p \in S. \tag{2.8}$$

The series converges normally for $0 \leqq r \leqq r_0 < \rho(f)$, $p \in S$. Hence $\hat{f}$ is a harmonic function defined on the ball with center zero and radius $\rho(f)$ in $\mathbb{R}^3$. At least if $\rho(f) > 1$, then $\hat{f}$ is the solution of the Dirichlet problem to find the harmonic function on the unit ball which is equal to f on the sphere S. Now $\rho(f)$ can be interpreted as the largest radius of a ball with center zero so there exists a harmonic function on this ball extending f, see Schäfke [8, Thm. 1, p. 139]. If $\rho(f) > 1$ then f, being the restriction of the analytic function $\hat{f}$, is itself analytic on the real analytic manifold S. The converse is also true: if f is analytic on S then $\rho(f) > 1$. This statement could be proved by methods from the theory of partial differential equations; however, we will see that it is a simple corollary of our main theorem in § 4.

There is a close connection between Laplace series and Legendre series (see, e.g., Hobson [4, p. 342]) which is crucial in most investigations of the convergence properties of Laplace series. Let (2.2) be the Laplace series associated with a given continuous function f on S. Introducing polar coordinates in S

$$k(\theta, \varphi) = (\sin\theta \cos\varphi, \sin\theta \sin\varphi, \cos\theta) \tag{2.9}$$

we can rewrite (2.3) in the form

$$f_n(p) = \frac{2n+1}{4\pi} \int_0^{\pi} \int_0^{2\pi} P_n(\langle p, k(\theta, \varphi)\rangle)(f \circ k)(\theta, \varphi) \sin\theta \, d\varphi \, d\theta.$$

Now if we set $p = (0, 0, 1)$ then we obtain

$$f_n(0, 0, 1) = \frac{2n+1}{2} \int_0^{\pi} P_n(\cos\theta) \sin\theta \, g(\theta) \, d\theta \tag{2.10}$$

where

$$g(\theta) := \frac{1}{2\pi} \int_0^{2\pi} (f \circ k)(\theta, \varphi) \, d\varphi \tag{2.11}$$

is the mean value of $(f \circ k)(\theta, \cdot)$ on $[0, 2\pi]$. If we substitute $t = \cos\theta$, $g(\theta) = \tilde{g}(t)$, then (2.10) yields

$$f_n(0, 0, 1) = \frac{2n+1}{2} \int_{-1}^{1} P_n(t) \tilde{g}(t) \, dt. \tag{2.12}$$

Hence $f_n(0, 0, 1)$, $n = 0, 1, 2, \cdots$, is equal to the sequence of coefficients in Legendre series associated with $\tilde{g}$, i.e.,

$$\tilde{g}(s) \sim \sum_{n=0}^{\infty} f_n(0, 0, 1) P_n(s), \qquad s \in [-1, 1].$$

To study the Laplace series at points different from the north pole we use its orthogonal invariance. Let $\mathcal{O}$ be a (real) orthogonal 3-by-3 matrix. Then $f \in L^2(S)$ implies $f \circ \mathcal{O} \in L^2(S)$, and if f_n is a spherical surface harmonic of degree n then $f_n \circ \mathcal{O}$ is also one. It

follows easily that if (2.2) is the Laplace series associated with a given function $f \in L^2(S)$, then

$$(f \circ \mathcal{O})(p) \sim \sum_{n=0}^{\infty} (f_n \circ \mathcal{O})(p), \qquad p \in S,$$

is the Laplace series associated with $f \circ \mathcal{O}$. In particular, the harmonic radius is orthogonal invariant, i.e., $\rho(f \circ \mathcal{O}) = \rho(f)$.

3. The manifold T. In the introduction we defined a subset T of $\mathbb{C}^3$ by

$$T = \{(x, y, z) \in \mathbb{C}^3 : x^2 + y^2 + z^2 = 1\}.$$

We consider T as a two-dimensional complex manifold. If $(x_0, y_0, z_0) \in T$ and, for example, x_0 is nonzero then we may take (y, z) as a local coordinate system in a neighborhood of (x_0, y_0, z_0). It will be useful to express the condition $x^2 + y^2 + z^2 = 1$ defining T in terms of the real and imaginary parts of x, y, z. If we write

$$(x, y, z) = u + iv \quad \text{with } u, v \in \mathbb{R}^3 \tag{3.1}$$

then we have

$$x^2 + y^2 + z^2 = \langle u, u \rangle - \langle v, v \rangle + 2i\langle u, v \rangle$$

where $\langle\ ,\ \rangle$ again denotes the usual inner product in $\mathbb{R}^3$. Hence

$$(x, y, z) \in T \quad \text{if and only if } \langle u, v \rangle = 0 \quad \text{and} \quad \langle u, u \rangle - \langle v, v \rangle = 1. \tag{3.2}$$

Similarly, we have

$$|x|^2 + |y|^2 + |z|^2 = \langle u, u \rangle + \langle v, v \rangle. \tag{3.3}$$

The above relations prove immediately the following lemma.

LEMMA 3.4 *Let $(x, y, z) \in T$ and let $\mathcal{O}$ be a (real) orthogonal 3-by-3 matrix. Then the transformed vector $(\tilde{x}, \tilde{y}, \tilde{z}) := \mathcal{O}(x, y, z)$ (here we do not distinguish between row and column vectors) also belongs to T and $|\tilde{x}|^2 + |\tilde{y}|^2 + |\tilde{z}|^2 = |x|^2 + |y|^2 + |z|^2$.*

Using the map $k : \mathbb{C}^2 \to T$ defined by (2.9) we have the following important lemma.

LEMMA 3.5. *For $0 \leqq t < \infty$, the following sets are identical*

$$\{(x, y, z) \in T : |x|^2 + |y|^2 + |z|^2 = \cosh 2t\}$$

$$= \{\mathcal{O} \circ k(\theta, \varphi) : \mathcal{O} \text{ orthogonal}, \varphi \in \mathbb{R}, |\operatorname{Im} \theta| = t\}$$

$$= \{\mathcal{O} \circ k(it, 0) : \mathcal{O} \text{ orthogonal}, \det \mathcal{O} = 1\}.$$

Proof. (1) To show that the second set is contained in the first one, let $\varphi \in \mathbb{R}$, $\theta = \theta_1 \pm it$, $\theta_1 \in \mathbb{R}$, and set $(x, y, z) := k(\theta, \varphi)$. Then we calculate

$$\begin{aligned} |x|^2 + |y|^2 + |z|^2 &= |\sin \theta|^2 + |\cos \theta|^2 \\ &= \sin^2 \theta_1 \cosh^2 t + \cos^2 \theta_1 \sinh^2 t + \cos^2 \theta_1 \cosh^2 t + \sin^2 \theta_1 \sinh^2 t \\ &= \cosh^2 t + \sinh^2 t = \cosh 2t. \end{aligned}$$

This shows that (x, y, z) is in the first set. Lemma 3.4 now implies that also $\mathcal{O}(x, y, z)$ is in this set for all orthogonal $\mathcal{O}$.

(2) To show that the first set is contained in the third one, let $(x, y, z) \in T$ and $|x|^2+|y|^2+|z|^2 = \cosh 2t$. We write (x, y, z) in the form (3.1). Then, by (3.2) and (3.3), $\langle u, u\rangle - \langle v, v\rangle = 1$ and $\langle u, u\rangle + \langle v, v\rangle = \cosh 2t$. Hence $\langle u, u\rangle = \cosh^2 t$ and $\langle v, v\rangle = \sinh^2 t$. By (3.2), u and v are orthogonal. Hence there exist two orthogonal unit vectors $\tilde{u}, \tilde{v} \in \mathbb{R}^3$ such that $u = \cosh t\, \tilde{u}$ and $v = \sinh t\, \tilde{v}$. Let $\mathcal{O}$ be the orthogonal matrix with $\det \mathcal{O} = 1$, first column $\tilde{v}$, and third column $\tilde{u}$. Then $(x, y, z) = u + iv = \mathcal{O}(i \sinh t, 0, \cosh t) = \mathcal{O} \circ k(it, 0)$, which proves that (x, y, z) is in the third set.

(3) Since the third set is obviously contained in the second one, the proof is complete. □

Let us now consider the sets

$$T_\gamma = \{(x, y, z) \in T: |x|^2+|y|^2+|z|^2 < \cosh 2\gamma\}$$

for $0 < \gamma \leqq \infty$ where $T_\gamma = T$ if $\gamma = \infty$. It is obvious that T_γ is an open subset of T. By Lemma 3.5, T_γ can be written in the form

$$T_\gamma = \{\mathcal{O} \circ k(it, 0): \mathcal{O} \text{ orthogonal, } \det \mathcal{O} = 1, 0 \leqq t < \gamma\}$$

which shows that T_γ is pathwise connected because the set of orthogonal 3-by-3 matrices with determinant 1 is pathwise connected. Consequently, T_γ is a domain in T. Lemma 3.5 also shows that the closure of T_γ and the boundary of T_γ with respect to T are given by

$$\bar{T}_\gamma = \{(x, y, z) \in T: |x|^2+|y|^2+|z|^2 \leqq \cosh 2\gamma\},$$

$$\partial T_\gamma = \{(x, y, z) \in T: |x|^2+|y|^2+|z|^2 = \cosh 2\gamma\}.$$

We again remark that

$$\bar{T}_0 := S \subset T_\beta \subset T_\gamma \subset T = T_\infty \quad \text{if } 0 < \beta < \gamma < \infty.$$

Every function holomorphic inside an ellipse with foci at the points ± 1 is uniquely determined by its values on the focal line $[-1, 1]$. Analogously, we have the following lemma.

LEMMA 3.6. *Let f, g be two functions holomorphic on T_γ such that $f = g$ on S. Then it follows that $f = g$ on T_γ.*

Proof. Let q_0 be any given point in T_γ. By Lemma 3.5, we can write

$$q_0 = \mathcal{O} \circ k(\theta_0, \varphi_0), \qquad \varphi_0 \in \mathbb{R}, \quad |\operatorname{Im} \theta_0| < \gamma.$$

Again by Lemma 3.5, we see that $f \circ \mathcal{O} \circ k(\theta, \varphi_0)$ and $g \circ \mathcal{O} \circ k(\theta, \varphi_0)$ are well defined and holomorphic functions of θ on the strip $|\operatorname{Im} \theta| < \gamma$. Since $f = g$ on S these functions agree on $\mathbb{R}$. The ordinary identity theorem now yields

$$f \circ \mathcal{O} \circ k(\theta, \varphi_0) = g \circ \mathcal{O} \circ k(\theta, \varphi_0) \quad \text{if } |\operatorname{Im} \theta| < \gamma.$$

Setting $\theta = \theta_0$ this proves $f(q_0) = g(q_0)$. □

4. The Laplace series associated with functions on T. Let f_n be any spherical surface harmonic of degree n. Since f_n is the restriction of a polynomial onto the unit sphere S, we see that f_n admits a holomorphic extension on the manifold T. This extension is uniquely determined by Lemma 3.6. In the rest of this paper a spherical surface harmonic will always be considered as a function defined on T.

Now let (2.2) be the Laplace series associated with a function $f \in L^2(S)$. Then, by (2.3), the spherical surface harmonic f_n is given by

$$f_n(x, y, z) = \frac{2n+1}{4\pi} \int_S P_n(xx'+yy'+zz') f(x', y', z')\, d(x', y', z') \tag{4.1}$$

for all $(x, y, z) \in S$. The right-hand side of (4.1) is a polynomial in x, y, z. Hence it represents a holomorphic function on T which is equal to f_n on S. It follows that (4.1) is true for all $(x, y, z) \in T$. We will now use (4.1) to prove an inequality that will play the same role for Laplace series as inequality (1.3) played for Legendre series.

PROPOSITION 4.2. *Let f_n be any spherical surface harmonic of degree n. Then*

$$|f_n(q)| \leqq (2n+1)\, e^{n\alpha} \|f_n | S\|_2 \quad \text{for all } q \in \bar{T}_\alpha.$$

Proof. We first prove the above inequality for $q = k(it, 0) = (i \sinh t, 0, \cosh t)$, where $0 \leqq t \leqq \alpha$. Then (4.1) with $f = f_n$ gives

$$f_n(q) = \frac{2n+1}{4\pi} \int_S P_n(z' \cosh t + ix' \sinh t) f_n(x', y', z')\, d(x', y', z').$$

The argument of the Legendre polynomial P_n lies in the set $\bar{E}_\alpha = \{\cos\theta : |\operatorname{Im}\theta| \leqq \alpha\}$ because $0 \leqq t \leqq \alpha$ and $(x', y', z') \in S$. Hence the above equation and (1.3) yield

$$|f_n \circ k(it, 0)| \leqq (2n+1) \|P_n | \bar{E}_\alpha\|_\infty \|f_n | S\|_2 \leqq (2n+1)\, e^{n\alpha} \|f_n | S\|_2. \tag{4.3}$$

If $q \in \bar{T}_\alpha$ is arbitrary then, by Lemma 3.5, we can find an orthogonal matrix $\mathcal{O}$ and $0 \leqq t \leqq \alpha$ such that $q = \mathcal{O} \circ k(it, 0)$. We then apply inequality (4.3) with $f_n \circ \mathcal{O}$ in place of f_n which proves the desired result. □

We now generalize inequality (1.6).

PROPOSITION 4.4. *Let f be a function holomorphic on T_γ where $0 < \gamma \leqq \infty$, and let (2.2) be the Laplace series associated with f. Then, for $0 < \beta < \gamma$ and all $p \in S$, the following inequality holds*

$$|f_n(p)| \leqq (2n+1) \frac{\pi}{2} \coth \beta\, e^{-n\beta} \|f | \bar{T}_\beta\|_\infty.$$

Proof. It is sufficient to prove the above inequality for $p = (0, 0, 1)$ because of the orthogonal invariance of Laplace series and Lemma 3.4. To estimate $|f_n(0, 0, 1)|$ we use (2.12), i.e., the connection between Laplace series and Legendre series. Since f is holomorphic on T_γ, Lemma 3.5 shows that the function g defined by (2.11) is holomorphic on the strip $|\operatorname{Im}\theta| < \gamma$. Clearly, g is an even function of period 2π which implies that the function $\tilde{g}$ defined by $\tilde{g}(\cos\theta) = g(\theta)$ is well defined and holomorphic on the domain E_γ. Now (2.12) and (1.6) give

$$|f_n(0, 0, 1)| \leqq (2n+1) \frac{\pi}{2} \coth \beta\, e^{-n\beta} \|\tilde{g} | \bar{E}_\beta\|_\infty. \tag{4.5}$$

It follows from (2.11) and Lemma 3.5 that, for $|\operatorname{Im}\theta| \leqq \beta$,

$$|\tilde{g}(\cos\theta)| = |g(\theta)| \leqq \max\{|f \circ k(\theta, \varphi)| : \varphi \in \mathbb{R}\} \leqq \|f | \bar{T}_\beta\|_\infty. \tag{4.6}$$

Equations (4.5) and (4.6) together prove the desired inequality. □

We are now in a position to prove our main theorem.

THEOREM 4.7. *Let f be a function holomorphic on T_γ where $0 < \gamma \leqq \infty$. Then the Laplace series associated with f converges normally on compact subsets of T_γ to the sum f. In particular, the series is uniformly convergent on compact subsets of T_γ.*

Proof. For given $0 \leqq \alpha < \gamma$ choose some β such that $\alpha < \beta < \gamma$. Then, by Propositions 4.2 and 4.4, the Laplace series associated with f has the following convergent majorant on $\bar{T}_\alpha$

$$\frac{\pi}{2} \coth \beta \|f | \bar{T}_\beta\|_\infty \sum_{n=0}^{\infty} (2n+1)^2\, e^{n(\alpha-\beta)}. \tag{4.8}$$

This proves that the Laplace series is normally convergent on compact subsets of T_γ. The sum of the series is holomorphic on T_γ and, by the completeness of the spherical surface harmonics in $L^2(S)$, it agrees on S with f. Hence the sum of the Laplace series is equal to f on T_γ because of Lemma 3.6. □

Let us note two simple corollaries of the above theorem.

COROLLARY 4.9. *Let $f \in L^2(S)$ and $0 < \gamma \leqq \infty$. Then the following statements are equivalent.*

(1) *The harmonic radius $\rho(f)$ of f defined by* (2.6) *or* (2.7) *is greater than or equal to e^γ.*

(2) *f admits a holomorphic extension onto T_γ.*

(3) *The Laplace series associated with f converges uniformly on compact subsets of T_γ.*

(4) *The Laplace series associated with f converges normally on compact subsets of T_γ.*

Proof. Obviously, (4) implies (3). If we assume (3), then the sum of Laplace series is holomorphic on T_γ and, as in the proof of Theorem 4.7, we see that this sum extends f. Hence (2) holds. Statement (2) implies (1) by Proposition 4.4. Finally, if we assume (1), then we prove (4) as in the proof of Theorem 4.7 replacing Proposition 4.4 by (1). □

We now prove a result announced in § 2.

COROLLARY 4.10. *Let $f \in L^2(S)$. Then f is analytic on S if and only if its harmonic radius $\rho(f)$ is larger than* 1.

Proof. If $\rho(f) > 1$, then f is analytic by the implication (1) $\Rightarrow$ (2) of Corollary 4.9.

Now let f be analytic on S. Then a standard compactness argument shows that there is a positive γ such that f can be continued holomorphically on T_γ. Hence $\rho(f) \geqq e^\gamma > 1$ by the implication (2) $\Rightarrow$ (1) of Corollary 4.9. □

For later applications, the following variant of Theorem 4.7 is useful. Since the spherical surface harmonics of degree n form a $(2n+1)$-dimensional subspace of $L^2(S)$, we can choose an orthonormal basis f_n^m, $m = -n, \cdots, n$, of this subspace. Then the Laplace series associated with a given function $f \in L^2(S)$ can be written as a double series

$$f(p) \sim \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \langle f, f_n^m \rangle_S f_n^m(p). \tag{4.11}$$

Concerning this series we have the following theorem.

THEOREM 4.12. *Let f be holomorphic on T_γ where $0 < \gamma < \infty$. Then the Laplace series* (4.11) *considered as a double series converges normally on compact subsets of T_γ to the sum f. More precisely, we have the estimate*

$$|\langle f, f_n^m \rangle_S f_n^m(q)| \leqq \frac{\pi}{2} \coth \beta \, \|f|\bar{T}_\beta\|_\infty (2n+1)^2 \, e^{n(\alpha-\beta)}$$

for $0 \leqq \alpha < \beta < \gamma$ and $q \in \bar{T}_\alpha$.

Proof. By Proposition 4.2, we have

$$|f_n^m(q)| \leqq (2n+1) \, e^{n\alpha} \quad \text{for } q \in \bar{T}_\alpha.$$

Hence the inequality of Cauchy-Schwarz yields

$$|\langle f, f_n^m \rangle_S f_n^m(q)| = |\langle f_n, f_n^m \rangle_S f_n^m(q)| \leqq (2n+1) \, e^{n\alpha} \|f_n|S\|_2.$$

This yields the desired inequality if we use Proposition 4.4. □

We conclude this section with the following remark. There is a generalization of Theorem 1.8 which states that every function holomorphic on the ring-shaped domain $E_\beta \setminus \bar{E}_\alpha$, $0 \leqq \alpha < \beta \leqq \infty$, can be expanded in a series in terms of Legendre polynomials

P_n and Legendre functions Q_n of the second kind, see, e.g., [5, Thm. 2, p. 256]. This expansion is related to that of Theorem 1.8 as Laurent's expansion is related to Taylor's expansion. We could try to find a similar expansion of Laurent's type valid for functions holomorphic on the domain $T_\beta \setminus \bar{T}_\alpha$, $0 \leqq \alpha < \beta \leqq \infty$. However, such an expansion does not yield anything new because every function holomorphic on $T_\beta \setminus \bar{T}_\alpha$ is already the restriction of a holomorphic function on T_β. To prove this we use Hartog's theorem [3, p. 18] which states that every function holomorphic on $G \setminus K$ can be continued holomorphically to a function on G provided G is a domain in $\mathbb{C}^n$ with $n \geqq 2$, K is a compact subset of G, and $G \setminus K$ is connected. The theorem remains true if G is a Stein manifold of dimension $n \geqq 2$, see [1, p. 144, Remark (2)]. We can apply this theorem because $G = T_\beta$ is a Stein manifold, $K = \bar{T}_\alpha$ is compact, and $T_\beta \setminus \bar{T}_\alpha$ is connected. The connectedness of $T_\beta \setminus \bar{T}_\alpha$ is a simple consequence of item 3 of Lemma 3.5.

5. Closing remarks. Theorem 4.12 can be used to prove a theorem on the expansion of a holomorphic function of two variables in a series of products of Lamé polynomials. In the following we will merely state this result without proof. The proof is contained in [11].

Consider Lamé's differential equation in its algebraic form [2, § 15.2, Eqn. (7)]

$$E'' + \frac{1}{2}\left(\frac{1}{u} + \frac{1}{u-1} + \frac{1}{u-a}\right) E' + \frac{ah - n(n+1)u}{4u(u-1)(u-a)} E = 0, \tag{5.1}$$

where $a > 1$ is fixed and n, h are parameters. It is well known that, for every even nonnegative integer n and every $m = 0, \cdots, n/2$, there is a uniquely determined value of the parameter h such that (5.1) admits a solution E which is a polynomial in u of degree $n/2$ having m zeros in the open interval $]0, 1[$ and $n/2 - m$ zeros in the open interval $]1, a[$. The polynomial E is called a Lamé polynomial and is denoted by E_n^m. It is uniquely determined by n and m up to a constant factor. We normalize the Lamé polynomials such that

$$\int_0^1 \int_1^a w(s, t) |E_n^m(s) E_n^m(t)|^2 \, dt \, ds = 1,$$

where

$$w(s, t) = \frac{1}{2\pi} \frac{t - s}{\sqrt{s(1-s)(a-s)}\sqrt{t(t-1)(a-t)}}, \qquad 0 < s < 1 < t < a.$$

The products $E_n^m(s) E_n^m(t)$ of Lamé polynomials are the expressions of special spherical surface harmonics of degree n in suitably chosen sphero-conal coordinates. It is therefore possible to apply Theorem 4.12 to obtain the following theorem.

THEOREM 5.2. *Let $g(s, t)$ be a function of two complex variables s, t which is holomorphic on the domain*

$$G_\gamma = \left\{ (s, t) \in \mathbb{C}^2 : \frac{|s||t|}{a} + \frac{|s-1||t-1|}{a-1} + \frac{|s-a||t-a|}{a(a-1)} < \cosh 2\gamma \right\},$$

where $0 < \gamma \leqq \infty$ ($G_\gamma = \mathbb{C}^2$ if $\gamma = \infty$). We assume that g is symmetric, i.e., *$g(s, t) = g(t, s)$ for all $(s, t) \in G_\gamma$. Then g can be expanded in the series*

$$g(s, t) = \sum_{\substack{n=0 \\ n \text{ even}}}^{\infty} \sum_{m=0}^{n/2} \left(\int_0^1 \int_1^a w(\sigma, \tau) g(\sigma, \tau) \overline{E_n^m(\sigma) E_n^m(\tau)} \, d\tau \, d\sigma \right) E_n^m(s) E_n^m(t),$$

which converges normally on compact subsets of G_γ.

This theorem on the expansion of a holomorphic function of two complex variables in a series of Lamé polynomials can be considered as an analogue to Neumann's theorem on the expansion of a holomorphic function of one variable in a series of Legendre polynomials.

REFERENCES

[1] H. BEHNKE AND P. THULLEN, *Theorie der Funktionen mehrerer komplexer Veränderlichen*, Springer-Verlag, Berlin, New York, 1970.

[2] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F. G. TRICOMI, *Higher Transcendental Functions*, III, McGraw-Hill, New York, 1955.

[3] G. M. HENKIN AND J. LEITERER, *Theory of Functions on Complex Manifolds*, Birkhäuser, Basel, 1984.

[4] E. W. HOBSON, *The Theory of Spherical and Ellipsoidal Harmonics*, Chelsea, New York, 1931.

[5] J. MEIXNER AND F. W. SCHÄFKE, *Mathieusche Funktionen und Sphäroidfunktionen*, Springer-Verlag, Berlin, New York, 1954.

[6] K. NEUMANN, *Über die Entwicklung einer Funktion nach den Kugelfunktionen*, Halle, 1862.

[7] G. SANSONE, *Orthogonal Functions*, Interscience, New York, 1959.

[8] F. W. SCHÄFKE, *Einführung in die Theorie der speziellen Funktionen der mathematischen Physik*, Springer, Berlin, New York, 1963.

[9] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1939.

[10] O. VOLK, *Über die Entwicklung von Funktionen zweier komplexen Veränderlichen nach Laméschen Funktionen*, Math. Z., 23 (1925), pp. 224-237.

[11] H. VOLKMER, *The expansion of a holomorphic function in a series of Lamé products*, preprint.

[12] E. T. WHITTAKER AND G.N. WATSON, *A Course of Modern Analysis*, Cambridge University Press, Cambridge, 1927.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 782-792, May 1990
© 1990 Society for Industrial and Applied Mathematics
013

A NOTE ON THE RELATIONSHIP BETWEEN STOKES MULTIPLIERS AND FORMAL SOLUTIONS OF ANALYTIC DIFFERENTIAL EQUATIONS*

G. K. IMMINK†

Abstract. This paper generalizes a result of Balser, Jurkat, and Lutz [*J. Math. Anal. Appl.*, 71 (1979), pp. 48-94] concerning the relation between the Stokes multipliers of a homogeneous linear differential equation with an irregular singularity and the asymptotic behaviour of the coefficients of the formal solutions of the equation.

Key words. irregular singularity, formal solutions, Stokes multipliers, Mellin transform, Cauchy-Heine transform, difference equation

AMS(MOS) subject classifications. 34A30, 34E05

0. Introduction. We consider homogeneous linear differential equations of a complex variable x with an irregular singularity at ∞. If (D) is such an equation of order $m \in \mathbb{N}$, it possesses m linearly independent formal solutions of the form

$$\hat{f}_j(t) = \hat{h}_j(t) t^{\rho} \exp q_j(t), \qquad j \in \{1, \cdots, m\}, \tag{0.1}$$

where $t = x^{1/p}$ for some $p \in \mathbb{N}$, $\hat{h}_j \in \mathbb{C}[\![t^{-1}]\!][\log t]$, $\rho_j \in \mathbb{C}$, and $q_j \in \mathbb{C}[t]$ for all $j \in \{1, \cdots, m\}$.

In [1], Balser, Jurkat, and Lutz establish a relation between the Stokes multipliers of a particular set of solutions and the asymptotic behaviour of the coefficients of the formal series $\hat{h}_j$ $(j = 1, \cdots, m)$, for second-order equations of unit rank. Schäfke (cf. [8]) has derived similar results for a class of first-order differential systems. In this note, these results are generalized to equations of arbitrary order and rank. We use Mellin transforms of solutions of an associated differential equation to represent the coefficients of $\hat{h}_j$ $(j \in \{1, \cdots, m\})$, and Cauchy-Heine transforms to represent solutions of (D) defined below. Our approach is based on the work of Ramis [6] and Duval [4].

1. A preliminary result. Let (D) be a differential equation of the type mentioned in the Introduction, of order m. In this section we introduce a particular system of solutions $\{f_j^{\nu}, j \in \{1, \cdots, m\}, \nu \in \mathbb{Z}\}$ of (D), with the property that, for each $j \in \{1, \cdots, m\}$ and each $\nu \in \mathbb{Z}$, f_j^{ν} is represented asymptotically by $\hat{f}_j$ as $t \to \infty$ in a certain sector S_{ν}. Other systems of solutions could be used instead, such as that discussed in [5, Satz IV', p. 99]; this discussion would lead to analogous results. The system defined below has a small technical advantage (cf. the remark at the end of § 2).

We begin by introducing some notation. By (D_t) we will denote the equation into which (D) is carried by the change of variable $t = x^{1/p}$. (D_t) possesses m formal solutions of the form (0.1). For all $i, j \in \{1, \cdots, m\}$ we will write

$$\rho_i - \rho_j = \rho_{ij}, \quad q_i - q_j = q_{ij}, \quad \deg q_{ij} = k(i, j).$$

We will assume that $k(i, j) \neq 0$ if $i \neq j$, in which case $\hat{h}_j \in \mathbb{C}[\![t^{-1}]\!]$ for all $j \in \{1, \cdots, m\}$. Without being essential, this restriction simplifies the argument presented below.

* Received by the editors March 23, 1988; accepted for publication (in revised form) April 21, 1989.
† Institute of Econometrics, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands.

If $i \neq j$, the leading term of q_{ij} is of the form

$$\lambda_{ij} t^{k(i,j)}, \qquad \lambda_{ij} \in \mathbb{C}^*$$

and we define

$$\theta_{ij}^{l} = \frac{1}{k(i,j)} \{\arg \lambda_{ij} - (2\lambda + 1)\pi\}, \qquad l \in \mathbb{Z}$$

for some fixed determination of $\arg \lambda_{ij}$. Let J denote the set of all ordered pairs (i, j) with $i, j \in \{1, \cdots, m\}$ and $i \neq j$. To each triplet $(i, j, l) \in J \times \mathbb{Z}$ we will assign an integer $\nu = n(i, j, l)$ and will write

$$\theta_{ij}^{l} = \theta_\nu \quad \text{and} \quad k(i,j) = k_\nu.$$

We choose $n(i, j, l)$ in such a way that, first, the mapping $n: J \times \mathbb{Z} \to \mathbb{Z}$ is a bijection, and second,

$$\theta_{\nu+1} \leqq \theta_\nu \quad \text{for all } \nu \in \mathbb{Z} \text{ and } k_{\nu+1} \leqq k_\nu \quad \text{whenever } \theta_\nu = \theta_{\nu+1}.$$

Let

$$N = \sum_{(i,j) \in J} k(i,j).$$

Noting that

$$\theta_{ij}^{l+k(i,j)} = \theta_{ij}^{l} - 2\pi,$$

we readily verify that

$$\theta_{\nu+N} = \theta_\nu - 2\pi. \tag{1.1}$$

We impose the following additional conditions on the mapping n:

$$n(i, j, l + k(i,j)) = n(i, j, l) + N. \tag{1.2}$$

Furthermore, we define

$$\sigma(i,j) = \{n(i,j,l) : l \in \mathbb{Z}\}, \qquad (i,j) \in J,$$

$$\sigma(j) = \bigcup_{i \in \{1,\cdots,m\} \setminus \{j\}} \sigma(i,j), \qquad j \in \{1, \cdots, m\},$$

$$J_\nu = \{(i,j) \in J : k(i,j) = k_\nu \text{ and } \theta_{ij}^{l} = \theta_\nu \text{ for some } l \in \mathbb{Z}\}, \qquad \nu \in \mathbb{Z}.$$

Note that $\nu \in \sigma(i,j)$ implies $(i,j) \in J_\nu$ but that the inverse is not true, as several θ_ν may coincide.

The directions $\arg t = -\theta_\nu - (\pi/2k_\nu)$, $\nu \in \mathbb{Z}$, of the Riemann surface of $\log t$, are the so-called Stokes directions of (D_t). The Stokes directions of (D) are given by $\arg x = -p(\theta_\nu + (\pi/2k_\nu))$, $\nu \in \mathbb{Z}$.

Let α and β be real numbers such that $\alpha < \beta$, and let $\mathbb{C}_\infty$ denote the Riemann surface of $\log t$. By $S(\alpha, \beta)$ we denote the sector

$$S(\alpha, \beta) = \{t \in \mathbb{C}_\infty : \alpha < \arg t < \beta\}.$$

For every $\nu \in \mathbb{Z}$ we define a sector $S_\nu = S(\alpha_\nu, \beta_\nu)$, where

$$\alpha_\nu = -\min_{\nu' < \nu} \left(\theta_{\nu'} + \frac{\pi}{2k_{\nu'}}\right), \qquad \beta_\nu = -\max_{\nu' \geqq \nu} \left(\theta_{\nu'} - \frac{\pi}{2k_{\nu'}}\right). \tag{1.3}$$

Obviously, $\alpha_\nu \leqq \alpha_{\nu+1}$ and $\beta_\nu \leqq \beta_{\nu+1}$. Furthermore, $\alpha_{\nu+N} = \alpha_\nu + 2\pi$ and $\beta_{\nu+N} = \beta_\nu + 2\pi$; hence

$$S_{\nu+N} = e^{2\pi i} S_\nu, \qquad \nu \in \mathbb{Z}.$$

Putting

$$\max_{(i,j)\in J} k(i,j)=\kappa_1,$$

we have

$$\theta_\nu+\frac{\pi}{2\kappa_1}\leqq \min_{\nu'<\nu+1}\left(\theta_{\nu'}+\frac{\pi}{2k_{\nu'}}\right)\leqq \theta_\nu+\frac{\pi}{2k_\nu},$$

$$\theta_\nu-\frac{\pi}{2k_\nu}\leqq \max_{\nu'\geqq\nu}\left(\theta_{\nu'}-\frac{\pi}{2k_{\nu'}}\right)\leqq \theta_\nu-\frac{\pi}{2\kappa_1}.$$

Hence it follows that

$$S\left(-\theta_\nu-\frac{\pi}{2\kappa_1},-\theta_\nu+\frac{\pi}{2\kappa_1}\right)\subset S_\nu\cap S_{\nu+1}\subset S\left(-\theta_\nu-\frac{\pi}{2k_\nu},-\theta_\nu+\frac{\pi}{2k_\nu}\right). \tag{1.4}$$

Any single mth order differential equation is equivalent to a system of first-order differential equations. Let $[D_t]$ denote the system of first-order differential equations corresponding to (D_t). This system possesses a formal fundamental matrix of the form

$$\hat{F}(t)=\hat{H}(t)t^R\exp Q(t), \tag{1.5}$$

where $\hat{H}\in Gl(m;\mathbb{C}[\![t^{-1}]\!])$, $R=\operatorname{diag}\{\rho_1,\cdots,\rho_m\}$, and $Q=\operatorname{diag}\{q_1,\cdots,q_m\}$. We will use the following result, which is an immediate consequence of Theorem 2.1 of [7].

THEOREM 1.6. *Equation $[D_t]$ possesses a system of fundamental matrices $\{\tilde{F}_\nu,\ \nu\in\mathbb{Z}\}$ with the following properties*:

(i) $\tilde{F}_\nu(t)\sim\hat{F}(t)$ *as* $t\to\infty$ *in* S_ν;

(ii) $\tilde{F}_{\nu+N}(t)=\tilde{F}_\nu(te^{-2\pi i})\exp 2\pi iR$;

(iii) $(\tilde{F}_\nu^{-1}\tilde{F}_{\nu+1}-I)_{ij}\neq 0$ $(i,j\in\{1,\cdots,m\})$ *implies* $(i,j)\in J_\nu$.

Proof. Let $\kappa_1>\kappa_2>\cdots>\kappa_r$ be the different values of $k(i,j)$, $(i,j)\in J$. According to Theorem 2.1 of [7], the matrix $\hat{H}$ in (1.5) can be factorized in the following way:

$$\hat{H}=\hat{H}_1\cdots\hat{H}_r,$$

where $\hat{H}_j\in Gl(m;\mathbb{C}\{t^{-1}\}_{s_j})$, $s_j=1+1/\kappa_j$. (For an explanation of the notation used in this proof we refer the reader to [6].) Let H_j^ν denote the matrix function obtained by analytic continuation on the Riemann surface of $\log t$ of the sum of $\hat{H}_j$ in a direction $\theta\in(\eta_j^\nu,\theta_j^\nu)$, where

$$\eta_j^\nu=\max\{\theta_{\nu'}:\nu'\geqq\nu,\ k_{\nu'}=\kappa_j\},\qquad \theta_j^\nu=\min\{\theta_{\nu'}:\nu'<\nu,\ k_{\nu'}=\kappa_j,\ \theta_{\nu'}>\eta_j^\nu\},$$

and let

$$\tilde{F}_\nu(t)=H_1^\nu(t)\cdots H_r^\nu(t)t^R\exp Q(t).$$

By Theorem 2.1 of [7], $\tilde{F}_\nu$ is a fundamental system of $[D_t]$. Furthermore, $H_j^\nu(t)\sim\hat{H}_j$ as $t\to\infty$ in

$$S_{j,\nu}\equiv S\left(-\theta_j^\nu-\frac{\pi}{2\kappa_j},-\eta_j^\nu+\frac{\pi}{2\kappa_j}\right).$$

Consequently, $\tilde{F}_\nu(t)\sim\hat{F}(t)$ as $t\to\infty$ in $\bigcap_{j=1}^r S_{j,\nu}$. Obviously, $S_\nu\subset S_{j,\nu}$ for all $j\in\{1,\cdots,r\}$ and thus (i) is satisfied. Due to (1.1), $H_j^{\nu+N}(t)=H_j^\nu(te^{-2\pi i})$ for all $j\in\{1,\cdots,r\}$ and hence it follows that (ii) holds as well. Now let $l\in\{1,\cdots,r\}$ such that $\kappa_l=k_\nu$. For all $j\neq l$ we have

$$\eta_j^{\nu+1}=\eta_j^\nu\leqq\theta_\nu\leqq\theta_j^{\nu+1}=\theta_j^\nu \tag{1.7}$$

and hence $H_j^{\nu+1} = H_j^{\nu}$. If $J_\nu = J_{\nu+1}$ and $\theta_\nu = \theta_{\nu+1}$, then (1.7) is also true for $j = l$, so we find

$$\tilde{F}_\nu^{-1}\tilde{F}_{\nu+1} = I \quad \text{if } J_\nu = J_{\nu+1} \text{ and } \theta_\nu = \theta_{\nu+1}.$$

Now suppose that $J_\nu \neq J_{\nu+1}$, or $\theta_\nu > \theta_{\nu+1}$. We have

$$\tilde{F}_\nu^{-1}\tilde{F}_{\nu+1} = \exp\{-Q(t)\}t^{-R}(H_r^\nu)^{-1}\cdots(H_l^\nu)^{-1}H_l^{\nu+1}H_{l+1}^\nu\cdots H_r^\nu t^R \exp Q(t).$$

Furthermore,

$$\eta_l^\nu = \theta_\nu = \theta_l^{\nu+1} \tag{1.8}$$

in this case. From (1.7) and (1.8) we deduce that

$$S_{l,\nu} \cap S_{l,\nu+1} = S\left(-\theta_\nu - \frac{\pi}{2k_\nu}, -\theta_\nu + \frac{\pi}{2k_\nu}\right) \subset S_{j,\nu} \quad \text{for all } j > l.$$

Now, $H_j^\nu \in Gl(m; A_{s_j}(S_{j,\nu}))$ for all $j \in \{1, \cdots, r\}$ and all $\nu \in \mathbb{Z}$, and thus $(H_l^\nu)^{-1}H_l^{\nu+1} - I \in \mathrm{End}\,(m; A_{0,s_l}(S_{l,\nu} \cap S_{l,\nu+1}))$. It follows that

$$(H_r^{(\nu)})^{-1}\cdots(H_l^\nu)^{-1}H_l^{\nu+1}\cdots H_r^\nu - I \in \mathrm{End}\,(m; A_{0,s_l}(S_{l,\nu} \cap S_{l,\nu+1})).$$

Consequently, for all h, $i \in \{1, \cdots, m\}$ we have

$$(\tilde{F}_\nu^{-1}\tilde{F}_{\nu+1} - I)_{hi} \exp q_{hi} \in A_{0,s_l}\left(S\left(-\theta_\nu - \frac{\pi}{2k_\nu}, -\theta_\nu + \frac{\pi}{2k_\nu}\right)\right).$$

This implies that either $(\tilde{F}_\nu^{-1}\tilde{F}_{\nu+1} - I)_{hi} = 0$, or $\exp q_{hi} \in A_{0,s_l}(S(-\theta_\nu - \pi/2k_\nu, -\theta_\nu + \pi/2k_\nu))$. The second possibility occurs only if $k(h,i) = k_\nu$ and $\mathrm{Re}\,\lambda_{hi}t^{k(h,i)} < 0$ for all $t \in S(-\theta_\nu - \pi/2k_\nu, -\theta_\nu + \pi/2k_\nu)$, i.e., if $(h,i) \in J_\nu$. This completes the proof of Theorem 1.6.

COROLLARY 1.9. *Equation* (D_t) *possesses a system of solutions* $\{f_j^\nu, j \in \{1, \cdots, m\}, \nu \in \mathbb{Z}\}$ *with the following properties*:

(i) $f_j^\nu(t) \sim \hat{f}_j(t)$ *as* $t \to \infty$ *in* S_ν;

(ii) $f_j^{\nu+N}(t) = f_j^\nu(t\,e^{-2\pi i})\,e^{2\pi i\rho_j}$;

(iii) *There exist complex numbers* s_ν, $\nu \in \mathbb{Z}$, *such that*

$$f_j^{\nu+1} - f_j^\nu = \begin{cases} s_\nu f_i^\nu & \text{if } \nu \in \sigma(i,j), \\ 0 & \text{otherwise.} \end{cases}$$

We will call the numbers s_ν the Stokes multipliers of the system of solutions $\{f_j^\nu, j \in \{1, \cdots, m\}, \nu \in \mathbb{Z}\}$. Note that (ii) and (iii) imply that $s_{\nu+N} = s_\nu\,e^{-2\pi i\rho_{ij}}$.

Proof of Corollary 1.9. We will prove the equivalent statement for the system $[D_t]$, i.e., we will prove the existence of fundamental matrices F_ν, $\nu \in \mathbb{Z}$, with the following properties:

(i)′ $F_\nu(t) \sim \hat{F}(t)$ as $t \to \infty$ in S_ν;

(ii)′ $F_{\nu+N}(t) = F_\nu(t\,e^{-2\pi i})\exp 2\pi i R$;

(iii)′ $(F_\nu^{-1}F_{\nu+1} - I)_{ij} \neq 0$ implies that $i \neq j$ and $\nu \in \sigma(i,j)$.

Let $\mu \in \mathbb{Z}$ and suppose that $J_{\mu-1} \neq J_\mu = \cdots = J_{\mu+r-1} \neq J_{\mu+r}$ and $\theta_\mu = \theta_{\mu+1} = \cdots \theta_{\mu+r-1}$ for some $r \in \mathbb{N}$. Thus J_μ consists of r pairs (i_h, j_h), $h = 1, \cdots, r$, with $k(i_h, j_h) = k_\mu$, and there exist integers l_h such that

$$\theta_{i_h j_h}^{l_h} = \theta_{\mu+h-1} = \theta_\mu, \qquad h \in \{1, \cdots, r\}. \tag{1.10}$$

Consequently, $\alpha_{\mu+1} = \alpha_{\mu+h}$ and $\beta_\mu = \beta_{\mu+h-1}$ for all $h \in \{1, \cdots, r\}$ (cf. (1.3)). If $r > 1$ it follows that

$$S_{\mu+h} = S_\mu \cap S_{\mu+r}, \qquad h \in \{1, \cdots, r-1\}. \tag{1.11}$$

We take $F_\mu = \tilde{F}_\mu$ and $F_{\mu+r} = \tilde{F}_{\mu+r}$. If $r=1$, (i)'-(iii)' are automatically satisfied for $\nu = \mu$. Now suppose that $r > 1$. Let

$$\tilde{F}_\mu^{-1} \tilde{F}_{\mu+r} = C.$$

According to Theorem 1.6, $(C-I)_{ij} = 0$ unless $(i,j) \in J_\mu$, i.e., unless $(i,j) \in \{(i_1,j_1), \cdots, (i_r,j_r)\}$. We readily verify that J_μ is an antisymmetric and transitive set (cf. [5]), i.e., $(i,j) \in J_\mu$ implies $(j,i) \notin J_\mu$, and if J_μ contains both (i,j) and (j,k) then it also contains (i,k). Due to these properties, C can be written uniquely as a product $C = C_1 \cdots C_r$, such that $(C_h - I)_{ij} = 0$ unless $(i,j) = (i_h, j_h)$, and thus $\mu + h - 1 \in \sigma(i,j)$ (cf. [5, p. 82]). Hence, if we choose

$$F_{\mu+h} = F_\mu C_1 \cdots C_h, \qquad h \in \{1, \cdots, r-1\},$$

then (iii)' obviously holds for $\nu = \mu + h - 1$, $h \in \{1, \cdots, r\}$, and hence for all $\nu \in \mathbb{Z}$. Furthermore, for all $h \in \{1, \cdots, r\}$ and $i, j \in \{1, \cdots, m\}$ we have

$$(F_{\mu+h} - F_{\mu+h-1})_{ij} = \begin{cases} 0 & \text{if } j \neq j_h, \\ (F_{\mu+h-1})_{ii_h} (C_h)_{i_h j_h} & \text{if } j = j_h. \end{cases} \tag{1.12}$$

Now suppose that $F_{\mu+h-1}(t) \sim \hat{F}(t)$ as $t \to \infty$ in $S_{\mu+h-1}$. Thus

$$(F_{\mu+h-1})_{ii_h}(t) \sim \hat{H}_{ii_h}(t) t^{\rho_{ih}} \exp q_{i_h}(t), \qquad t \to \infty \text{ in } S_{\mu+h-1}. \tag{1.13}$$

From (1.10) we infer that $\operatorname{Re} \lambda_{i_h j_h} t^{k(i_h, j_h)} < 0$ for all $t \in S(-\theta_\mu - (\pi/2k_\mu), -\theta_\mu + (\pi/2k_\mu))$, and hence, in view of (1.4) and (1.11), for all $t \in S_{\mu+h}$, provided $h < r$. This implies that $\exp q_{i_h j_h}(t)$ decreases exponentially as $t \to \infty$ in $S_{\mu+h}$. With (1.13) it follows that

$$(F_{\mu+h-1})_{ii_h}(t) \exp(-q_{j_h}(t)) \sim 0 \quad \text{as } t \to \infty \text{ in } S_{\mu+h}.$$

Combining this with (1.12), we conclude that $F_{\mu+h}(t) \sim \hat{F}(t)$ as $t \to \infty$ in $S_{\mu+h}$. By means of induction on h this property can be established for all $h \in \{0, \cdots, r-1\}$ and thus (i)' is true for all $\nu \in \mathbb{Z}$. As $J_\nu = J_{\nu+N}$ for all $\nu \in \mathbb{Z}$, we have

$$F_{\mu+N} = \tilde{F}_{\mu+N}, \qquad F_{\mu+r+N} = \tilde{F}_{\mu+r+N}$$

if μ is chosen as before. Let

$$\tilde{F}_{\mu+N}^{-1} \tilde{F}_{\mu+r+N} = \tilde{C}.$$

It follows from Theorem 1.6, property (ii), that

$$\tilde{C} = \exp(-2\pi i R) \tilde{F}_\mu^{-1} \tilde{F}_{\mu+r} \exp 2\pi i R. \tag{1.14}$$

Furthermore, $\tilde{C}$ can be written uniquely as a product $\tilde{C} = \tilde{C}_1 \cdots \tilde{C}_r$ such that $(\tilde{C}_h - I)_{ij} = 0$ unless $\mu + h + N - 1 \in \sigma(i,j)$, or equivalently (due to (1.2)), $\mu + h - 1 \in \sigma(i,j)$. With (1.14) it follows that

$$\tilde{C}_h = \exp(-2\pi i R) C_h \exp 2\pi i R, \qquad h \in \{1, \cdots, r\}.$$

Thus, $F_{\mu+h+N}(t) = \tilde{F}_{\mu+N}(t) \exp(-2\pi i R) C_1 \cdots C_h \exp 2\pi i R$, and hence, in view of Theorem 1.6, property (ii),

$$F_{\mu+h+N}(t) = \tilde{F}_\mu(t e^{-2\pi i}) C_1 \cdots C_h \exp 2\pi i R = F_{\mu+h}(t e^{-2\pi i}) \exp 2\pi i R, \qquad h \in \{1, \cdots, r\}.$$

This proves (ii)' for all $\nu \in \mathbb{Z}$.

2. An integral representation for the coefficients of $\hat{h}_j$. Let $j \in \{1, \cdots, m\}$. If $\hat{f}_j$ is a formal solution of (D_t) of the form (0.1), then (under the assumption made in § 1)

$\hat{h}_j$ is a formal power series solution of the equation (D_t^j) resulting from a change of the unknown y in (D_t) to z, where

$$z(t) = y(t)t^{-\rho_j} \exp(-q_j(t)).$$

Throughout this section it will be assumed that (D) has no singularities in the finite complex plane but at most a regular one at the origin. In that case, both (D_t) and (D_t^j) have at most a regular singularity at the origin of the complex t-plane. (D_t^j) may be written in the following form:

$$D_t^j z \equiv \sum_{l=0}^{m} \sum_{h=0}^{N_j} a_{hl} t^h \left(t\frac{d}{dt}\right)^l z = 0,$$

where $N_j = \sum_{i\in\{1,\cdots,m\}: i\neq j} k(i,j)$, $a_{hl} \in \mathbb{C}$, $a_{0m} \neq 0$, and $a_{N_j 1} \neq 0$. Let

$$\hat{h}_j = \sum_{n=0}^{\infty} \hat{h}_{jn} t^{-n}, \qquad \hat{h}_{j0} = 1.$$

The coefficients $\hat{h}_{jn}$ with $n \geqq 1$ can be determined by means of a recursive relation, or equivalently, by solving an N_jth-order difference equation (Δ^j) of the form

$$(\Delta^j) \qquad \sum_{h=0}^{N_j} \sum_{l=0}^{m} a_{hl}(-1)^l (n+h)^l y_{n+h} = 0,$$

subject to N_j initial conditions.

For each $\nu \in \mathbb{Z}$ let ψ_ν denote the function defined by

$$(2.1) \qquad \psi_\nu(t) = f_i^\nu(t) t^{-\rho_j} \exp(-q_j(t)) \quad \text{if } \nu \in \sigma(i,j),$$

where f_i^ν is one of the solutions of (D_t) mentioned in Corollary 1.9. Obviously, ψ_ν is a solution of (D_t^j), represented asymptotically by

$$\hat{h}_i(t) t^{\rho_{ij}} \exp q_{ij}(t)$$

as $t \to \infty$ in S_ν. Let γ_ν be a half-line in $S_\nu \cap S_{\nu+1}$ starting from zero. Due to (1.4),

$$\operatorname{Re} \lambda_{ij} t^{k(i,j)} < 0 \quad \text{for all } t \in S_\nu \cap S_{\nu+1}, \qquad \nu \in \sigma(i,j).$$

Therefore, ψ_ν decreases exponentially as $t \to \infty$ on γ_ν. Furthermore, since the origin is at most a regular singular point of (D_t^j), $\psi_\nu(1/t)$ has at most polynomial growth as $t \to 0$. Hence the integral

$$(2.2) \qquad \hat{h}^\nu(n) \equiv -\frac{1}{2\pi i} \int_{\gamma_\nu} \psi_\nu(t) t^{n-1}\, dt$$

exists for sufficiently large n. With the use of partial integration it is easily verified that the function $\hat{h}^\nu$ defined by (2.2) satisfies the difference equation (Δ^j) if $\nu \in \sigma(j)$ and n is larger than some integer n_0. Let

$$\tilde{\sigma}(j) = \sigma(j) \cap \{1, \cdots, N\}.$$

For each $j \in \{1, \cdots, m\}$ the N_j functions $\hat{h}^\nu$ with $\nu \in \tilde{\sigma}(j)$ are linearly independent (cf. the remark below) and, consequently, form a fundamental system of solutions of (Δ^j). Hence there exist complex numbers c_ν, $\nu \in \{1, \cdots, N\}$ such that, for each $j \in \{1, \cdots, m\}$,

$$(2.3) \qquad \hat{h}_{jn} = \sum_{\nu \in \tilde{\sigma}(j)} c_\nu \hat{h}^\nu(n), \qquad n \geqq n_0.$$

Remark. Let $i, j \in J$ and $l \in \mathbb{Z}$ such that $\nu \equiv n(i, j, l) \in \{1, \cdots, N\}$. With the aid of the saddle-point method it can be shown that $\hat{h}^\nu$ admits an asymptotic representation of the form

$$\hat{h}^\nu(n) \sim C_\nu \Gamma\left(\frac{n+\rho_{ij}}{k_\nu}\right) \exp\left(p_\nu(n+\rho_{ij})\right)(1+g_\nu(n)), \qquad n \to \infty, \tag{2.4}$$

where p_ν is a polynomial in n^{1/k_ν} of degree not exceeding k_ν, without constant term, $C_\nu \in \mathbb{C}^*$, and $g_\nu \in n^{-1/k_\nu}\mathbb{C}[\![n^{-1/k_\nu}]\!]$. p_ν and C_ν are completely determined by q_{ij} and l and, conversely, p_ν determines q_{ij} and l (cf. [2], [3]). (Due to (1.4), the sector $S_\nu \cap S_{\nu+1}$ contains a saddle point of the function $q_{ij}(t) + n \log t$ if n is sufficiently large, and (2.4) can be proved by a straightforward application of the saddle-point method. For other systems of solutions, such as that discussed in [5], the proof of (2.4) is slightly more involved, as it requires a study of the asymptotic behaviour of ψ_ν outside S_ν.)

The assumption in § 1 that $k(h, i) \neq 0$ for all $(h, i) \in J$ implies that $q_{hj} \neq q_{ij}$ if $h \neq i$. Therefore, if $j \in \{1, \cdots, m\}$ is given, q_{ij} and l determine ν. Now suppose that ν_1, $\nu_2 \in \tilde{\sigma}(j)$ and $\nu_1 \neq \nu_2$. Then it follows that $p_{\nu_1} \neq p_{\nu_2}$ and (2.4) shows that $\hat{h}^{\nu_1}$ and $\hat{h}^{\nu_2}$ must be linearly independent.

3. A relation between the Stokes multipliers s_ν and the coefficients of $\hat{h}_j$. We begin by assuming, as we did in the previous section, that (D) has no singularities in the finite complex plane but at most a regular one at the origin. Under this condition we have the following generalization of the result of Balser, Jurkat, and Lutz [1] mentioned in the Introduction.

THEOREM 3.1. *For all $\nu \in \{1, \cdots, N\}$ the coefficients c_ν in* (2.3) *are equal to the Stokes multipliers s_ν defined in Corollary* 1.9.

Proof. Consider the function h^ν defined by

$$h^\nu(t) = t^{-n_0+1} \int_{\gamma_\nu} \frac{\psi_\nu(\tau)\tau^{n_0-1}}{2\pi i(\tau - t)}\, d\tau, \qquad \nu \in \{1, \cdots, N\},$$

where n_0, γ_ν, and ψ_ν have been defined in the previous section. (The function $h^\nu(t)t^{n_0-1}$ is a Cauchy-Heine transform of $\psi_\nu(t)t^{n_0-1}$.) h^ν is analytic in $\mathbb{C}\backslash \mathrm{Im}\ \gamma_\nu$. As γ_ν is an arbitrary half-line in $S_\nu \cap S_{\nu+1} = S(\alpha_{\nu+1}, \beta_\nu)$ (cf. (1.3)), starting from zero, h^ν can be continued analytically to the sector

$$\mathbf{S}_\nu \equiv S(\alpha_{\nu+1}, \beta_\nu + 2\pi)$$

by continuously changing the direction of γ_ν. It follows from (2.2) and Proposition 4.2 of [6] that h^ν admits the asymptotic representation

$$h^\nu(t) \sim \sum_{n=n_0}^{\infty} \hat{h}^\nu(n) t^{-n}, \qquad t \to \infty \text{ in } \mathbf{S}_\nu \tag{3.2}$$

(and even that h^ν is Gevrey of order $1 + 1/k_\nu$). Furthermore, it is easily seen that

$$h^\nu(t) - h^\nu(t\, e^{2\pi i}) = \psi_\nu(t), \qquad t \in S_\nu \cap S_{\nu+1}. \tag{3.3}$$

For each $j \in \{1, \cdots, m\}$ let h_j^ν be defined by

$$h_j^\nu(t) = \sum_{n=0}^{n_0-1} \hat{h}_{jn} t^{-n} + \sum_{\mu \in \tilde{\sigma}(j): \mu < \nu} s_\mu h^\mu(t) + \sum_{\mu \in \tilde{\sigma}(j): \mu \geqq \nu} s_\mu h^\mu(t\, e^{2\pi i}),$$

where s_μ, $\mu = 1, \cdots, N$, are the Stokes multipliers mentioned in Corollary 1.9. From (3.2) and (3.3) we conclude that h_j^ν is analytic on $\mathbb{C}_\infty$ and admits the following

asymptotic representation:

$$h_j^\nu(t) \sim \sum_{n=0}^{n_0-1} \hat{h}_{jn} t^{-n} + \sum_{n=n_0}^{\infty} \sum_{\mu \in \tilde{\sigma}(j)} s_\mu \hat{h}^\mu(n) t^{-n} \tag{3.4}$$

as $t \to \infty$ in

$$\bigcap_{\mu \in \tilde{\sigma}(j): \mu < \nu} \mathbf{S}_\mu \bigcap_{\mu \in \tilde{\sigma}(j): \mu \cong \nu} e^{-2\pi i} \mathbf{S}_\mu = S_\nu.$$

Moreover, we have

$$h_j^{\nu+1}(t) - h_j^\nu(t) = \begin{cases} 0 & \text{if } \nu \notin \tilde{\sigma}(j), \quad \nu < N, \\ s_\nu(h^\nu(t) - h^\nu(t e^{2\pi i})) = s_\nu \psi_\nu(t) & \text{if } \nu \in \tilde{\sigma}(j), \quad \nu < N, \end{cases}$$

$$h_j^1(t e^{-2\pi i}) - h_j^N(t) = \begin{cases} 0 & \text{if } N \notin \sigma(j), \\ s_N(h^N(t) - h^N(t e^{2\pi i})) = s_N \psi_N(t) & \text{if } N \in \sigma(j). \end{cases}$$

Putting

$$h_j^\nu(t) t^{\rho_j} \exp q_j(t) = g_j^\nu(t),$$

and using (2.1) and Corollary 1.9, we find the following relations:

$$g_j^{\nu+1} - g_j^\nu = f_j^{\nu+1} - f_j^\nu, \qquad \nu = 1, \cdots, N-1,$$
$$g_j^1(t e^{-2\pi i}) e^{2\pi i \rho_j} - g_j^N(t) = f_j^1(t e^{-2\pi i}) e^{2\pi i \rho_j} - f_j^N(t),$$

or equivalently,

$$\begin{aligned} &g_j^{\nu+1} - f_j^{\nu+1} = g_j^\nu - f_j^\nu, \qquad \nu = 1, \cdots, N-1, \\ &\{g_j^1(t e^{-2\pi i}) - f_j^1(t e^{-2\pi i})\} e^{2\pi i \rho_j} = g_j^N(t) - f_j^N(t). \end{aligned} \tag{3.5}$$

These identities show that $(g_j^1(t) - f_j^1(t)) t^{-\rho_j}$ is analytic in $\mathbb{C} \setminus \{0\}$. From the asymptotic properties of g_j^ν and f_j^ν given by (3.4) and Corollary 1.9, in combination with (3.5), we deduce that

$$(g_j^1(t) - f_j^1(t)) t^{-\rho_j} \exp(-q_j(t)) \sim \sum_{n=n_0}^{\infty} a_j(n) t^{-n}, \qquad t \to \infty \text{ in } \mathbb{C},$$

where $a_j(n) = \sum_{\nu \in \tilde{\sigma}(j)} s_\nu \hat{h}^\nu(n) - \hat{h}_{jn} = \sum_{\nu \in \tilde{\sigma}(j)} (s_\nu - c_\nu) \hat{h}^\nu(n)$ (cf. (2.3)). Consequently, $\sum_{n=n_0}^{\infty} a_j(n) t^{-n}$ is a convergent power series and its sum is $(g_j^1(t) - f_j^1(t)) t^{-\rho_j} \exp(-q_j(t))$. In view of (2.4) this implies that $s_\nu - c_\nu = 0$ for all $\nu \in \tilde{\sigma}(j)$ (and hence $g_j^\nu = f_j^\nu$ for all $\nu \in \{1, \cdots, N\}$).

If (D) has an irregular singularity at the origin, or other finite singular points, the argument presented above is no longer valid. For example, the integral on the right-hand side of (2.2) may not exist. We therefore modify the path γ_ν in the following way. Let $t_\nu \in S_\nu \cap S_{\nu+1}$, such that all finite singularities of (D_t) are contained within the disk $\{t \in \mathbb{C}: |t| < |t_\nu|\}$ and let $\tilde{\gamma}_\nu$ be a half-line in $S_\nu \cap S_{\nu+1}$, starting from t_ν. Now the integral

$$\hat{h}^\nu(n) \equiv -\frac{1}{2\pi i} \int_{\tilde{\gamma}_\nu} \psi_\nu(t) t^{n-1} \, dt \tag{3.6}$$

is well defined for all $n > 0$ and admits the asymptotic representation (2.4) (which is independent of t_ν). Furthermore, let

$$h^\nu(t) = t \int_{\tilde{\gamma}_\nu} \frac{\psi_\nu(\tau) \tau^{-1}}{2\pi i(\tau - t)} \, d\tau.$$

h^ν is analytic in the "sector" $\tilde{\mathbf{S}}_\nu = \{t \in \mathbf{S}_\nu : |t| > |t_\nu|\}$. According to Proposition 4.2 of [6], h^ν is Gevrey of order $1+1/k_\nu$ and is represented asymptotically by

$$\sum_{n=0}^{\infty} \hat{h}^\nu(n) t^{-n}$$

as $t \to \infty$ in $\mathbf{S}_\nu$. Moreover, we have

$$h^\nu(t) - h^\nu(t e^{2\pi i}) = \psi_\nu(t), \quad t \in S_\nu \cap S_{\nu+1}, \quad |t| > |t_\nu|.$$

Defining, for each $j \in \{1, \cdots, m\}$, a function h_j^ν by

$$h_j^\nu(t) = \sum_{\mu \in \tilde{\sigma}(j): \mu < \nu} s_\mu h^\mu(t) + \sum_{\mu \in \tilde{\sigma}(j): \mu \geqq \nu} s_\mu h^\mu(t e^{2\pi i})$$

and proceeding as in the proof of Theorem 3.1, we conclude that the function a_j^ν defined by

$$a_j^\nu(t) \equiv f_j^\nu(t) t^{-\rho_j} \exp(-q_j(t)) - h_j^\nu(t), \quad j \in \{1, \cdots, m\}, \quad \nu \in \{1, \cdots, N\},$$

is holomorphic at ∞ and independent of ν. Thus we finally obtain the following result.

THEOREM 3.7. *For all $j \in \{1, \cdots, m\}$ the coefficients of the formal power series $\hat{h}_j$ can be written in the form*

$$\hat{h}_{jn} = \sum_{\nu \in \tilde{\sigma}(j)} s_\nu \hat{h}^\nu(n) + a_{jn}, \qquad n \in \mathbb{N}, \tag{3.8}$$

where the numbers s_ν are the Stokes multipliers defined in Corollary 1.9, *the a_{jn} are complex numbers with the property that* $\limsup_{n\to\infty} |a_{jn}|^{1/n} < \infty$, *and $\hat{h}_\nu$ is given by* (3.6) *and represented asymptotically by* (2.4) *as $n \to \infty$.*

In general, for a given $j \in \{1, \cdots, m\}$, one of the functions $\hat{h}_\nu$ with $\nu \in \tilde{\sigma}(j)$ will dominate the others as $n \to \infty$. With the aid of (3.8) and (2.4), the corresponding Stokes multipliers s_ν can be determined from the asymptotic behaviour of $\hat{h}_{jn}$. In many cases we may even do a little better, due to the fact that the Stokes multipliers are analytic functions of certain coefficients of the equation. In particular, if $k(i, j)$ is constant for all $(i, j) \in J$ ("one-leveled" equation), it is possible to determine all Stokes multipliers, if the asymptotic behaviour of $\hat{h}_{jn}$ as a function of these coefficients is known. This is illustrated by the example in § 4.

4. An example. We illustrate the foregoing with a very simple example:

$$(D) \qquad x^2 y''(x) + \sigma x y'(x) = (x^4 + \alpha x^3 + \beta x^2 + \gamma x + \delta) y(x), \qquad \sigma, \alpha, \beta, \gamma, \delta \in \mathbb{C}.$$

This equation has two linearly independent formal solutions $\hat{f}_1$ and $\hat{f}_2$ of the form

$$\hat{f}_1(x) = \hat{h}_1(x) x^\rho \exp q(x), \qquad \hat{f}_2(x) = \hat{h}_2(x) x^{-\rho-\sigma-1} \exp(-q(x)),$$

where $q(x) = \frac{1}{2}(x^2 + \alpha x)$, $\rho = \frac{1}{2}(\beta - \frac{1}{4}\alpha^2 - \sigma - 1)$, and $\hat{h}_j(x) = 1 + \sum_{n=1}^{\infty} \hat{h}_{jn} x^{-n}$, $j = 1, 2$.

Using the notation introduced in the previous sections we have $p = 1$, $k(1, 2) = k(2, 1) = 2$, $\lambda_{12} = 1$, and $\lambda_{21} = -1$. Hence, choosing $\theta_0 = 0$, we find

$$\theta_\nu = -\frac{\nu\pi}{2}, \quad S_\nu = S\left(\frac{\nu\pi}{2} - \frac{3\pi}{4}, \frac{\nu\pi}{2} + \frac{\pi}{4}\right), \quad \nu \in \mathbb{Z}.$$

In this particular ("one-leveled") case, (D) possesses two unique solutions, f_1^1 and f_1^3, represented asymptotically by $\hat{f}_1$ as $x \to \infty$ in $S_1 \cup S_2$ and $S_3 \cup S_4$, respectively, and two unique solutions f_2^2 and f_2^4 represented asymptotically by $\hat{f}_2$ as $x \to \infty$ in $S_2 \cup S_3$ and $S_4 \cup S_5$, respectively.

The substitution

$$y(x) = z(x)x^{\rho} \exp q(x)$$

takes (D) into

$$(D^1) \qquad x^2 z''(x) + \{2x^3 + \alpha x^2 + (2\rho + \sigma)x\} z'(x) + \{(\alpha\rho + \tfrac{1}{2}\alpha\sigma - \gamma)x + \rho(\rho + \sigma - 1) - \delta\} z(x) = 0.$$

The coefficients $\hat{h}_{1n}$ satisfy the difference equation

$$(\Delta^1) \qquad 2(n+2)y_{n+2} + \left\{\alpha\left(n - \rho - \frac{\sigma}{2} + 1\right) + \gamma\right\} y_{n+1} - \{(n-\rho)(n-\rho-\sigma+1) - \delta\} y_n = 0$$

with initial conditions

$$y_0 = 1, \qquad y_1 = \tfrac{1}{2}\left(\alpha\rho + \alpha\frac{\sigma}{2} - \gamma\right). \tag{4.1}$$

For sufficiently large n, according to (2.2), (2.3) and Theorem 3.1, h_{1n} may be written in the form

$$\hat{h}_{1n} = -\frac{s_2}{2\pi i}\int_0^{\infty e^{\pi i}} \psi_2(x)x^{n-1}\,dx - \frac{s_4}{2\pi i}\int_0^{\infty e^{2\pi i}} \psi_4(x)x^{n-1}\,dx,$$

where $\psi_\nu(x) \equiv f_2^\nu(x)x^{-\rho}\exp(-q(x))$, $\nu = 2, 4$. The asymptotic behaviour of $\hat{h}_{1n}$ is given by

$$\hat{h}_{1n} = \frac{1}{4\pi i}\exp\left\{\frac{\alpha^2}{8} - \pi i(2\rho+\sigma)\right\}\Gamma\left(\frac{n-2\rho-\sigma-1}{2}\right) \cdot \left[s_2(-1)^n \exp\left(\alpha\sqrt{\frac{n}{2}}\right)(1+o(1)) - s_4 \exp\left\{-\pi i(2\rho+\sigma) - \alpha\sqrt{\frac{n}{2}}\right\}(1+o(1))\right], \quad n \to \infty. \tag{4.2}$$

Suppose, for example, that $\operatorname{Re}\alpha > 0$. Then it follows that

$$s_2 = 4\pi i \exp\left\{\pi i(2\rho+\sigma) - \frac{\alpha^2}{8}\right\} \lim_{n\to\infty} \hat{h}_{1n}\Gamma\left(\frac{n-2\rho-\sigma-1}{2}\right)^{-1} (-1)^n \exp\left(-\alpha\sqrt{\frac{n}{2}}\right). \tag{4.3}$$

If, on the other hand, $\operatorname{Re}\alpha < 0$, we have

$$s_4 = -4\pi i \exp\left\{2\pi i(2\rho+\sigma) - \frac{\alpha^2}{8}\right\} \lim_{n\to\infty} \hat{h}_{1n}\Gamma\left(\frac{n-2\rho-\sigma-1}{2}\right)^{-1} \exp\left(\alpha\sqrt{\frac{n}{2}}\right). \tag{4.4}$$

As both Stokes multipliers are entire functions of α, once their values for $\operatorname{Re}\alpha > 0$ or $\operatorname{Re}\alpha < 0$ are known, other values may be found by analytic continuation. Unfortunately, it is not easy to determine the Stokes multipliers as functions of the coefficients of the equation. In general, (4.3) and (4.4) will merely yield approximations of s_2 and s_4 for given values of these coefficients. (In this connection it might be worthwhile to study the parameter dependence of the solutions of the difference equation (Δ^1).) To conclude this section we mention two particular cases in which (Δ^1) can be solved explicitly.

(i) $\alpha = \gamma = 0$. Then we have

$$\hat{h}_{12n} = \frac{\Gamma(n-a)\Gamma(n-b)}{n!\Gamma(-a)\Gamma(-b)}, \quad \hat{h}_{12n+1} = 0, \quad n \in \mathbb{N},$$

where a and b are the roots of the equation $(2x-\rho)(2x-\rho-\sigma+1)-\delta=0$. With (4.2) we find

$$s_2 = s_4 \exp(-\pi i\beta) = -\frac{2\pi i \exp \pi i\beta}{\Gamma(-a)\Gamma(-b)}.$$

(ii) $\sigma=\gamma=\delta=0$. We readily verify that in this case $\hat{h}_{1n}$ may be represented by

$$\hat{h}_{1n} = \frac{1}{\sqrt{\pi}}\binom{n-\rho-1}{n}\int_{-\infty}^{\infty} e^{-(x+\alpha/2)^2}x^n\,dx. \tag{4.5}$$

Comparison of the asymptotic behaviour of the right-hand side of (4.5) with (4.2) yields the following expressions for the Stokes multipliers s_2 and s_4:

$$s_2 = -s_4 \exp(-2\pi i\rho) = \frac{i\sqrt{\pi}\,2^{-\rho}\exp(2\pi i\rho - \frac{1}{4}\alpha^2)}{\Gamma(-\rho)},$$

in agreement with the result found by Sibuya (cf. [9, Thm. 22.2]).

Remark. After this paper was completed I received a preprint by M. Loday in which she presents a method for computing the Birkhoff invariants of differential systems of order 2 (cf. [10]). Following her approach, we can improve the computation of the Stokes multipliers s_2 and s_4 in the preceding example by making the change of variable $x=\xi-\frac{1}{2}\alpha$. This essentially reduces the polynomial $q(x)$ to a single term $(\frac{1}{2}\xi^2)$ and makes the exponential factors $\exp(\alpha\sqrt{n/2})$ and $\exp(-\alpha\sqrt{n/2})$ disappear from the asymptotic representation (4.2). Consequently, both s_2 and s_4 can be computed from the asymptotic behaviour of the coefficients $\hat{h}_{1n}$ of the modified formal solution $\tilde{f}_1$, regardless of the value of α.

However, in the case of higher-order equations, the appearance of exponentials with different orders of growth in the asymptotic representations of the coefficients $\hat{h}_{jn}$, as in (4.2), cannot usually be avoided.

Acknowledgments. Thanks are due to Professors B. L. J. Braaksma and J. P. Ramis for their helpful comments.

REFERENCES

[1] W. Balser, W. B. Jurkat, and D. A. Lutz, *Birkhoff invariants and Stokes' multipliers for meromorphic linear differential equations*, J. Math. Anal. Appl., 71 (1979), pp. 48–94.

[2] N. G. de Bruyn, *Asymptotic Methods in Analysis*, North-Holland, Amsterdam, New York, 1961.

[3] A. Duval, *Etude asymptotique d'une intégrale analogue à la fonction "Γ modifiée,"* in *Equations différentielles et systèmes de Pfaff dans le champ complexe*—II, Lecture Notes in Math. 1015, Springer-Verlag, Berlin, New York, 1983, pp. 50–63.

[4] ———, *Equations aux différences algébriques: solutions méromorphes dans* $\mathbb{C}-\mathbb{R}^-$, *Système fondamental de solutions holomorphes dans un demi-plan*, Lecture Notes in Math. 1015, Springer-Verlag, Berlin, New York, 1983, pp. 102–135.

[5] W. B. Jurkat, *Meromorphe Differentialgleichungen*, Lecture Notes in Math. 637, Springer-Verlag, Berlin, New York, 1978.

[6] J. P. Ramis, *Les séries k-sommables et leurs applications*, Lecture Notes in Phys. 126, Springer-Verlag, Berlin, New York, 1980, pp. 178–199.

[7] ———, *Filtration Gevrey sur le groupe de Picard Vessiot d'une équation différentielle irrégulière*, Informes de Matematica Serie A-045/85 (1985).

[8] R. Schäfke, *Über das globale Verhalten der Normallösungen von* $x'(t)=(B+t^{-1}A)x(t)$ *und zweier Arten von assoziierten Funktionen*, Math. Nachr., 121 (1985), pp. 123–145.

[9] Y. Sibuya, *Global Theory of a Second Order Linear Ordinary Differential Equation with a Polynomial Coefficient*, North-Holland, Amsterdam, New York, 1975.

[10] M. Loday-Richaud, *Calcul des invariants de Birkhoff des systèmes d'ordre deux*, Preprint 88-31, Orsay, France, 1988.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 793-798, May 1990
© 1990 Society for Industrial and Applied Mathematics
014

ON THE CONNECTION PROBLEM FOR SOME SCHRÖDINGER EQUATIONS IN RELATION TO THE BICONFLUENT HEUN DIFFERENTIAL EQUATION*

B. LEAUTE,† G. MARCILHACY,† R. PONS,‡ AND J. SKINAZI‡

Abstract. Following a previous paper concerning the eigenvalue problem, greater detail is given for determining of the so-called connection coefficients for the Schrödinger equation of rotating harmonic, three-dimensional, anharmonic oscillators and for a class of confinement potentials related to the biconfluent Heun differential equation.

Key words. connection coefficients, difference equation

AMS(MOS) subject classification. 34B25

1. Introduction. Many physical problems concern eigenvalue problems. Such is the case for the rotating harmonic, three-dimensional, and doubly anharmonic oscillators and a class of confinement potentials [1]-[4]. In a previous paper [5] it has been shown that in these four cases, the corresponding Schrödinger equations are, in fact, now Fuschian differential equations of the biconfluent Heun (BCH) type [5]. This linear equation which has one regular singularity at the origin and one irregular singularity at $t=+\infty$ of the fourth species, is characterised in the Ince [6] classification by the formula $[0, 1, 1_4]$.

The study of eigenvalues for this equation is closely connected with the so-called two-connection problem. As a rule we have to determine the constant coefficients of the linear combinations between the two fundamental sets of solutions at the origin and at $t=+\infty$. It is a very difficult problem. In the neighbourhood of the origin the solutions are represented by convergent series, but at infinity by formal series. Moreover, when the variable is complex, the asymptotic representations of the solutions are sectorially dependent (Stokes phenomenon).

In this paper, by application to the BCH equation of the general-theoretical Kowalevskii [8]-[11] method, we present some results concerning this problem in a more tractable form.

Previously, this problem has been studied in great detail and with very interesting results by Kohno [12]-[15], in 1970 for an nth order linear differential equation with an irregular singularity of rank 2, and in 1974 for an nth-order linear differential equation with an irregular singularity of arbitrary rank.

2. The biconfluent Heun equation (BCH). Previously [5], it was shown that the four Schrödinger equations reduce to the BCH equation. This equation has been written in canonical form [7]. Nevertheless, here it is convenient to put the BCH in the following form:

$$t^2\varphi'' - t(a_{1,0}+a_{1,1}t+a_{1,2}t^2)\varphi' - (a_{2,0}+t^4)\varphi = 0, \tag{1}$$

where $a_{i,j}$ are four irreducible parameters. Equation (1) has two singular points: a regular singular point at zero, and an irregular singular point at infinity. By making

* Received by the editors May 11, 1988; accepted for publication (in revised form) May 26, 1989.

† U.R.A., Centre National de Recherche Scientifique 769, Université P. et M. Curie, Institut Henri Poincaré, 11, rue P. et M. Curie, 75231 Paris Cedex 05, France.

‡ Laboratoire de Physique Mathématique des Plasmas, Université Pierre et Marie Curie, 4, Place Jussieu, 75252 Paris Cedex 05, France.

the standard transformation

$$y = x^{\rho}\, e^{ax+(b/2)x^2}\varphi(kt)$$

on the canonical BCH equation [7]

$$xy''+(1+\alpha-\beta x-2x^2)y'+((\gamma-\alpha-2)x-\tfrac{1}{2}(\delta+(1+\alpha)\beta))=0, \tag{2}$$

we obtain (1), with the four relations:

$$\begin{aligned}
&-\tfrac{1}{2}[\delta+(1+\alpha)\beta]+a(1+\alpha)-\rho\beta+2a\rho=0,\\
&\gamma-\alpha-2-a\beta+\alpha\beta+a^2+2b+2\rho(b-1)=0,\\
&2(1-b)a+b\beta=0,\\
&b(b-2)k^4=1.
\end{aligned}$$

It follows that

$$\begin{aligned}
&a_{1,0}=-(2\rho+1+\alpha), \qquad a_{1,1}=k(2a-\beta),\\
&a_{1,2}=2k^2(1-b), \qquad a_{2,0}=-\rho(\rho+\alpha).
\end{aligned} \tag{3}$$

According to the analytic theory of differential equations, we can write the following results in the vicinity of the two singularities.

(a) Solution at the origin ($t=0$). At the origin it holds that a fundamental system of solutions is given in the form of convergent series

$$\varphi_j = t^{\rho_j}\sum_{s=0}^{\infty} G_j(s)t^s, \tag{4}$$

where $j=1,2$ and ρ_j is a root of the characteristic equation

$$E(\rho)\equiv\rho(\rho-1)-a_{1,0}\rho-a_{2,0}=0. \tag{5}$$

If ρ_1, ρ_2 are the roots of this equation, we suppose that $\rho_1\rho_2\neq 0$, $\rho_1\neq\rho_2$, and $\rho_1-\rho_2$ is a noninteger. The coefficients $G_j(s)$ satisfy the difference equation

$$\begin{aligned}
&[(\rho_j+s)(\rho_j+s-1)-a_{1,0}(\rho_j+s)-a_{2,0}]G_j(s)\\
&\quad-[(\rho_j+s-1)a_{1,1}]G_j(s-1)-[(\rho_j+s-2)a_{1,2}]G_j(s-2)-G(s-4)=0,
\end{aligned} \tag{6}$$

with the conditions

$$G_j(0)=1, \qquad G_j(-1)=G(-2)=G_j(-3)=0. \tag{7}$$

(b) Solution at infinity ($t=+\infty$). In any section S of the complex t-plane, defined by $|\operatorname{Arg} t|<(\pi/2)-\varepsilon$ there exists, when $t\to\infty$, a fundamental system of solutions of (1) [19]:

$$\varphi^{(k)}(t,S)\simeq t^{\mu_k}\, e^{Q_k(t)}\sum_{s=0}^{\infty} h^{(k)}(s)t^{-s}, \tag{8}$$

where $k=1,2$ and $Q_k(t)=\lambda_1^{(k)}t+\lambda_1^{(k)}t^2$. The quantities $\lambda_1^{(k)}$ are the roots, supposed nonzero and distinct, of the characteristic equation

$$F(\lambda_2)\equiv 4\lambda_2^2-2a_{1,2}\lambda_2-a_{2,4}=0, \tag{9}$$

with

$$\lambda_2^{(1)}=\frac{\sqrt{a_{1,2}^2+4}}{4}, \qquad \lambda_2^{(2)}=\frac{\sqrt{a_{1,2}^2-4}}{4}.$$

By direct substitution of (8) in (1) we find the other quantities $\lambda_1^{(k)}, \lambda_2^{(k)}, \mu_1, \mu_2$:

$$\lambda_1^{(1)} = \frac{2a_{1,1}\lambda_2^{(1)}}{4\lambda_2^{(1)} - a_{1,2}}, \qquad \lambda_1^{(2)} = \frac{2a_{1,1}\lambda_2^{(2)}}{4\lambda_2^{(2)} - a_{1,2}},$$

$$\mu_1 = \frac{a_{1,1}\lambda_1^{(1)} + 2a_{1,0}\lambda_2^{(1)} - \lambda_1^{(1)2} - 2\lambda_2^{(1)}}{4\lambda_2^{(1)} - a_{1,2}}, \tag{10}$$

$$\mu_2 = \frac{a_{1,1}\lambda_1^{(2)} + 2a_{1,0}\lambda_2^{(2)} - \lambda_1^{(2)2} - 2\lambda_2^{(2)}}{4\lambda_2^{(2)} - a_{1,2}}.$$

It follows that [14]

$$\lambda_1^{(1)} + \lambda_1^{(2)} = a_{1,1}, \qquad \mu_1 + \mu_2 = a_{1,0} - 1. \tag{11}$$

3. The connection coefficients. Between the solutions at the origin and at infinity of (1) there hold the formal relations

$$\varphi_j(t) = C_j^{(1)}(S)\varphi^{(1)}(t, S) + C_j^{(2)}(S)\varphi^{(2)}(t, S), \tag{12}$$

where S is some section of the complex plane with angle less than $\pi/2$.

Instead of determining directly the coefficients $C_j^{(k)}$, we strive to find the asymptotic behaviour of the solutions at the origin and to obtain a relation of type (12). The nonperiodic coefficients $A_j^{(\nu)}$ defined below, are called "connection coefficients" [10]. If we put

$$G_j(s) = \sum_{\nu=1}^{4} A_j^{(\nu)} f_j^{(\nu)}(s), \tag{13}$$

where $\{f_j^{(\nu)}(s)\}_{\nu=1}^4$ is a fundamental system of solutions to (6), then, taking (7) into account, we find that $A_j^{(\nu)}$ are solutions of

$$\sum_{\nu=1}^{4} A_j^{(\nu)} f_j^{(\nu)}(0) = 1, \qquad \sum_{\nu=1}^{4} A_j^{(\nu)} f_j^{(\nu)}(s) = 0, \tag{14}$$

for $s = -1, -2, -3$, where $j = 1, 2$. Now the functions $f_j^{(\nu)}(s)$ can be expressed [9], [12], [13] in terms of series as follows:

$$f_j^{(\nu)}(s) = \sum_{l=0}^{\infty} g_{j,r(\nu)}^{k(r)}(s+l) h^{k(\nu)}(l), \tag{15}$$

where the coefficients $h^{k(\nu)}(l)$ are given by (8), and

$$|\arg s| < \pi - \varepsilon, \quad 0 < \varepsilon < \pi, \quad \nu = 1, 2, 3, 4, \quad r(\nu) = E(\nu/2).$$

$E(z)$ is the integer part of the number z, and $k(\nu) = \nu + 1 - 2r(\nu)$. Besides, $g_{j,r}^k(z)$ in (15) is the following modified Γ function [16]:

$$g_{j,r}^k(z) = \frac{1}{2i\pi} \int_{L_r^{(k)}} \tau^{1/2[-z+(\mu_k-\rho_j)]-1} e^{\lambda_1^{(k)}\tau^{1/2} + \lambda_2^{(k)}\tau} \, d\tau, \tag{16}$$

where $z \in \mathbb{R}$.

The contour $L_r^{(k)}$ is composed of the three following parts ($1 \leqq |\tau| < \infty$):

(1) The ray $\arg(\tau) = -\pi(k+2r)$ starting at ∞;

(2) The circle $|\tau| = 1$ described on the negative direction;

(3) The ray $\arg(\tau) = \pi(k+2r) + 2\pi$ going to ∞.

The function $g_j^k(z)$ is an arbitrary solution of the hypergeometric difference equation [10], [12]:

$$(z + \rho_j - \mu_k) g_j^{(k)}(z) = \lambda_1^{(k)} g_j^{(k)}(z-1) + 2\lambda_2^{(k)} g_j^{(k)}(z-2). \tag{17}$$

This is demonstrated by integrating by parts. Explicitly, the functions $g_{j,r}^{(k)}(z)$ can be written

$$
\begin{aligned}
&(\nu=1) \qquad g_{j,0}^{(2)}(z)=\frac{1}{2i\pi}\int_{L_0^{(2)}} \tau^{1/2[-z+(\mu_2-\rho_j)]-1}\, e^{\lambda_1^{(2)}\tau^{1/2}+\lambda_2^{(2)}\tau}\, d\tau,\\
&(\nu=2) \qquad g_{j,1}^{(1)}(z)=\frac{1}{2i\pi}\int_{L_1^{(1)}} \tau^{1/2[-z+(\mu_1-\rho_j)]-1}\, e^{\lambda_1^{(1)}\tau^{1/2}+\lambda_2^{(1)}\tau}\, d\tau,\\
&(\nu=3) \qquad g_{j,1}^{(2)}(z)=\frac{1}{2i\pi}\int_{L_1^{(2)}} \tau^{1/2[-z+(\mu_2-\rho_j)]-1}\, e^{\lambda_1^{(2)}\tau^{1/2}+\lambda_2^{(2)}\tau}\, d\tau,\\
&(\nu=4) \qquad g_{j,2}^{(1)}(z)=\frac{1}{2i\pi}\int_{L_2^{(1)}} \tau^{1/2[-z+(\mu_1-\rho_j)]-1}\, e^{\lambda_1^{(1)}\tau^{1/2}+\lambda_2^{(1)}\tau}\, d\tau.
\end{aligned}
\tag{18}
$$

We infer the asymptotic behaviour of these functions $g_{j,r}^{(k)}(z)$ from the general result [8], [9]

$$
\begin{aligned}
g_{j,r(\nu)}^{k(\nu)}(2z+\alpha) \underset{z\to\infty}{\simeq} (k_\nu)^{z+\alpha/2}\, e^{[z-z\ln(z)+\alpha_1^{(\nu)}z]}\, z^{1/2[(\mu_k-\rho_j-1)-\alpha]}\\
*\left[\tilde{g}_{j,0}^{(\nu)}+\sum_{l=1}^{\infty}\tilde{g}_{j,l}^{(\nu)}(\alpha)z^{-l/2}\right],
\end{aligned}
\tag{19}
$$

where $\alpha\in\mathbb{C}$, $\kappa_\nu=|2\lambda_2^{k(\nu)}|\, e^{\pi\nu i}$ the constants $\alpha_1^{(\nu)}$ and $\tilde{g}_{j,0}^{(\nu)}$ are independent of α and $\tilde{g}_{j,0}^{(\nu)}\neq 0$.

Later we need only the asymptotic behaviour of the $f_j^{(\nu)}(s)$ functions defined by (15). This is

$$
f_j^{(\nu)}(s)=g_{j,r(\nu)}^{k(\nu)}(s)\{1+O(s^{-1/2})\}, \quad s\to\infty, \quad |\arg s|<\pi-\varepsilon,
\tag{20}
$$

$$
f_j^{(\nu)}(s+\alpha)=\left(\frac{\kappa_\nu}{s}\right)^{\alpha/2} g_{j,r(\nu)}^{k(\nu)}(s)\{1+O(s^{-1/2})\} \quad \text{if } \alpha\in\mathbb{C}.
\tag{21}
$$

4. Adjoint equation of (6). The coefficients of $\{A_j^{(\nu)}\}$ are solutions of system (14). Now, using results of Norlund [16], [17] on the so-called adjoint equation of the difference equation (6), we find [9] that it is possible to calculate the $\{A_j^{(\nu)}\}$ without solving the system (14).

The adjoint equation of (6) can be written [17]:

$$
\begin{aligned}
&[(\rho_j+s)(\rho_j+s-1)-a_{1,0}(\rho_j+s)-a_{2,0}]\mu_j(s-4)\\
&\quad -[(\rho_j+s)a_{1,1}]\mu_j(s-3)-[(\rho_j+s)a_{1,2}]\mu_j(s-2)-\mu_j(s)=0.
\end{aligned}
\tag{22}
$$

Now, simple relations exist between the solution of (22) and that of the previous equation (6).

Let $D_j(s)$ be the Casoratis determinant of the fundamental system of solutions of (6):

$$
\begin{aligned}
D_j(s)&=D_j[f_j^{(1)}(s),f_j^{(2)}(s),f_j^{(3)}(s),f_j^{(4)}(s)]\\
&\equiv\begin{vmatrix}
f_j^{(1)}(s) & f_j^{(2)}(s) & f_j^{(3)}(s) & f_j^{(4)}(s)\\
f_j^{(1)}(s+1) & f_j^{(2)}(s+1) & f_j^{(3)}(s+1) & f_j^{(4)}(s+1)\\
f_j^{(1)}(s+2) & f_j^{(2)}(s+2) & f_j^{(3)}(s+2) & f_j^{(4)}(s+2)\\
f_j^{(1)}(s+3) & f_j^{(2)}(s+3) & f_j^{(3)}(s+3) & f_j^{(4)}(s+3)
\end{vmatrix}.
\end{aligned}
\tag{23}
$$

$D_j^{(\nu)}$ is the algebraic complement of the νth element of the last line. Then a fundamental solution of (22) is given by

$$\mu_j^{(\nu)}(s)=\frac{D_j^{(\nu)}(s+1)}{p^j(s)D_j(s+1)},\qquad \nu=1,2,3,4, \tag{24}$$

where

$$p^j(s)=[(\rho_j+s+4)(\rho_j+s+3)-a_{1,0}(\rho_j+s+4)-a_{2,0}].$$

It follows that $D_j(s)$ given by (23) can be written:

$$D_j(s)=\left[\prod_{l=0}^{3} g_{j,r(l)}^{k(l)}(s)\right]\mathcal{V}_\alpha(1,i,-1,-i)(1+O(s^{-1/2})), \tag{25}$$

where $\mathcal{V}_\alpha(1,i,-1,-i)$ is the corresponding Vandermonde determinant. Likewise,

$$D_j^{(\nu)}(s)=\left[\prod_{\substack{l=0\\ l\neq\nu}}^{3} g_{j,r(l)}^{k(l)}(s)\right]\mathcal{W}_\nu(1+O(s^{-1/2})), \tag{26}$$

where

$$\mathcal{W}_1=\mathcal{V}_\alpha(i,-1,-i)=4i,\qquad \mathcal{W}_2=\mathcal{V}_\alpha(1,-1,-i)=4,$$
$$\mathcal{W}_3=\mathcal{V}_\alpha(1,i,-i)=-4i,\qquad \mathcal{W}_4=\mathcal{V}_\alpha(1,i,-1)=-4,$$

and we obtain for the asymptotic behaviour of the $\mu_j^{(\nu)}(s)$

$$\mu_j^{(\nu)}(s)=\frac{C_\nu}{p^j(s)g_{j,r(\nu)}^{k(\nu)}(s+1)}(1+O(s^{-1/2})),\qquad C_\nu=\frac{\mathcal{W}_\nu}{\mathcal{V}_\alpha(1,i,-1,-i)}, \tag{27}$$

where $s\to\infty$, $|\arg s|<\pi-\varepsilon$, $0<\varepsilon<\pi$. From (24) it follows that [17]

$$\sum_{\nu=0}^{3}\mu_j^{(\nu)}(s)f_j^{(\nu)}(s+l)=0,\qquad \sum_{\nu=0}^{3}p^j(s)\mu_j^{(\nu)}(s)f_j^{(\nu)}(s+3)=1, \tag{28}$$

where $j=1,2$ and $l=0,1,2$. Comparing the system (14) with the system (28), it may be deduced that the connection coefficients $A_j^{(\nu)}$ are given by

$$A_j^{(\nu)}=\mu_j^{(\nu)}(-3)p^j(-3)=\mu_j^{(\nu)}(-3)((\rho_j+1)\rho_j-a_{1,0}(\rho_j+1)-a_{2,0}), \tag{29}$$

where $\nu=0,1,2,3$ and $j=1,2$. These coefficients are therefore solutions of the adjoint equation (27) for $s=-3$ with the asymptotic conditions

$$\lim_{s\to\infty}(\mu_j^{(\nu)}(s)g_{j,r(\nu)}^{k(\nu)}(s+1)p^j(s))=C_\nu. \tag{30}$$

5. Conclusion. The two fundamental sets of solutions, at the origin and at infinity, are formally connected by (12). In this paper, instead of determining directly the coefficients involved in these relations, we have found an asymptotic representation of the convergent power series solutions at origin (4), of type (12). We have obtained these asymptotic behaviours by means of asymptotic set solutions of the adjoint linear difference equation (22), and have found that, in fact, the connection coefficients are expressed in terms of the values of particular solutions of that adjoint difference equation.

A previous paper [5] has shown how, by means of integral equations, it is possible as a rule to solve the eigenvalue problem for the Schrödinger equation of type $[0, 1, 1_4]$. However, it is not an easy task to solve this integral equation. In the present paper a very different way is proposed using a more classical method. This method can be used for other Schrödinger equations having only two singularities with one irregular singular at infinity.

REFERENCES

[1] D. MASSON, *The rotating harmonic oscillator eigenvalue problem, continued fractions and analytic continuation*, J. Math. Phys., 24 (1983), pp. 2074-2087.

[2] G. P. FLESSAS, *On the three-dimensional anharmonic oscillator*, Phys. Lett. A, 78 (1980), pp. 19-21.

[3] R. N. CHAUDHURI, *The Hill determinant: an application to a class of confinement potential*, J. Phys. A, 16 (1983), pp. 209-211.

[4] R. N. CHAUDHURI AND B. MUKHERJEE, *On the* $\mu x^2+\lambda x^4+\eta x^6$ *interaction*, J. Phys. A, 17 (1984), pp. 3327-3334.

[5] B. LÉAUTÉ AND G. MARCILHACY, *On the Schrödinger equation of rotating harmonic three-dimensional and doubly anharmonic oscillators and a class of confinement potentials in connection with the biconfluent Heun differential equation*, J. Phys. A, 19 (1986), pp. 3527-3533.

[6] E. L. INCE, *Ordinary Differential Equations*, Dover, New York, 1986.

[7] A. DECARREAU, P. MARONI, AND A. ROBERT, *Sur les équations confluentes de l'équation de Heun*, Ann. Soc. Sci. Bruxelles Ser. 1, 92 (1978), pp. 151-189.

[8] M. A. KOWALEVSKII, *Construction of the difference equation in order to determine the connection coefficients of the fundamental solution*, Vestnik Leningrad Univ. Math., 8 (1985), pp. 94-97. (In Russian.)

[9] ———, *Asymptotics of functions that generalize the Euler gamma function*, J. Soviet Math., 20 (1982), pp. 1826-1830.

[10] ———, *Construction of the Stokes multipliers for an equation with two singular points*, Vestnik Leningrad Univ. Math., 14 (1982), pp. 135-141.

[11] ———, *Determination of the connection between two fundamental families of solutions of a linear differential equation*, Vestnik Leningrad Univ. Math., 14 (1982), pp. 39-46.

[12] M. KOHNO, *A two point connection problem for the n-th order single linear ordinary differential equations with an irregular singular point of rank two*, Japan J. Math., 40 (1970), pp. 11-62.

[13] ———, *A two point connection problem for general linear ordinary differential equations*, Hiroshima Math. J., 4 (1974), pp. 293-338.

[14] ———, *Derivation of Stokes multipliers*, Hiroshima Math. J., 14 (1904), pp. 247-256.

[15] ———, *A two point connection problem*, Hiroshima Math. J., 9 (1979), pp. 61-135.

[16] N. G. DE BRUIJN, *Asymptotic Methods in Analysis*, North-Holland, Amsterdam, New York, 1958.

[17] N. E. NORLUND, *Leçons sur les équations linéaires aux différences finies*, Gauthier-Villars, Paris, 1929.

[18] A. A. MIROLIOUBOV AND M. A. SOLDATOV, *Les équations aux différences linéaires*, Moscow, 1981. (In Russian.)

[19] G. BIRKHOFF, *Formal theory of irregular linear difference equations*, Acta Math., 54 (1930), pp. 205-246.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 799-802, May 1990
© 1990 Society for Industrial and Applied Mathematics
015

ZEROS OF CHEBYSHEV POLYNOMIALS ASSOCIATED WITH A COMPACT SET IN THE PLANE*

E. B. SAFF† AND V. TOTIK‡

Abstract. It is proved that the zeros of the Chebyshev polynomials associated with a compact set in the plane having connected interior and complement stay away from the boundary if and only if the set is bounded by an analytic curve.

Key words. Chebyshev polynomials, zeros, analytic curve, Faber polynomials

AMS(MOS) subject classifications. 30C15, 41A50

Let K be an infinite compact subset of the complex plane $\mathbb{C}$. The unique nth-degree monic polynomial $T_n^K(z) = T_n(z) = z^n + \cdots$ with minimal supremum norm on K is called the nth *Chebyshev polynomial* associated with K. It is well known that the zeros of T_n^K lie in the convex hull of the set K. For the case when K is the unit disk, $T_n^K(z) = z^n$, $n = 0, 1, \cdots$, so it is possible for all the zeros of T_n^K to lie in the interior of K. The aim of this paper is to characterize those sets K for which the zeros stay away from the boundary of K.

Let G_∞ be the unbounded component of the complement $\mathbb{C}\setminus K$ of K. Obviously the Chebyshev polynomials associated with K are the same as those associated with $\mathbb{C}\setminus G_\infty$; therefore in what follows we will assume that $K = \mathbb{C}\setminus G_\infty$, i.e., the complement of K is connected. Widom [5] has proved that for every closed subset S of G_∞ there is a natural number n_S such that each T_n^K can have at most n_S zeros in S. Thus, most of the zeros are close to K. In the case where K has empty interior we actually know the asymptotic distribution of the zeros of T_n^K; namely, it coincides with the equilibrium measure of the set K (see [1]). This result is no longer true if K has nonempty interior, as the above-mentioned example of the unit disk shows. It seems to be a very difficult problem to determine the distribution of the zeros (if it exists at all) for general K's. In connection with this question our aim is to prove the following theorem.

THEOREM. *Let K be a compact subset of $\mathbb{C}$ with connected interior and complement. Then the zeros of the Chebyshev polynomials T_n^K stay away from the boundary of K if and only if K is bounded by an analytic curve.*

By "staying away from the boundary" we mean that for some neighborhood of the boundary there is no zero of T_n^K in this neighborhood for all large n. The proof shows that the same result holds if by "staying away from the boundary" we mean that for some neighborhood of the boundary there are at most $o(n)$ zeros of T_n^K in this neighborhood for $n \to \infty$.

By an analytic curve we mean a simple closed curve γ that has a parametric representation $\gamma_1(t) + i\gamma_2(t)$, $t \in [0, 2\pi]$, where γ_1 and γ_2 are analytic functions on $[0, 2\pi]$.

It seems likely that our result is valid in a somewhat more general form; namely, if K has disconnected interior, then the zeros stay away from the boundary exactly when K is bounded by a finite number of (in this case not necessarily simple) analytic

* Received by the editors December 7, 1988; accepted for publication April 26, 1989.

† Institute for Constructive Mathematics, Department of Mathematics, Tampa, Florida 33620. The research of this author was partially supported by National Science Foundation grant DMS-862-0098.

‡ Bolyai Institute Aradi V. tere 1, 6720 Szeged, Hungary. The research of this author was partially supported by Hungarian National Science Foundation for Research grant 1157.

curves. However, in this formulation "staying away" must mean the weaker $o(n)$ version discussed above as can be seen from the example: $K=\{z\,|\,|z^2+1|\leqq 1\}$. In fact, this K is bounded by an analytic (though not simple) curve, but the symmetry of K with respect to the origin implies that $T^K_{2n+1}(0)=0$ for all n.

Proof. (Sufficiency.) We need the Faber polynomials associated with the set K.

Our assumption is that K is bounded by a simple closed analytic curve γ. Thus the complement G_∞ of K in $\bar{\mathbb{C}}:=\mathbb{C}\cup\{\infty\}$ can be mapped conformally onto the exterior of a circle $C_R=\{w\,|\,|w|=R\}$ by a function φ normalized by $\varphi(\infty)=\infty$, $\lim_{z\to\infty}\varphi(z)/z=1$ (cf. [2, § 14]). Then R is the logarithmic capacity of K and since, without loss of generality, this may be assumed to be 1, in what follows we take $R=1$, i.e., φ maps G_∞ conformally onto the exterior of the unit disk. If

$$\varphi(z)=z+\alpha_0+\frac{\alpha_{-1}}{z}+\cdots$$

is the Laurent expansion of φ at infinity, then the expansion of φ^n is of the form

$$\varphi^n(z)=z^n+\alpha^{(n)}_{n-1}z^{n-1}+\cdots+\alpha^{(n)}_0+\frac{\alpha^{(n)}_{-1}}{z}+\cdots.$$

The polynomials

$$F_n(z):=z^n+\alpha^{(n)}_{n-1}z^{n-1}+\cdots+\alpha^{(n)}_0$$

are called the *Faber polynomials* of K.

Since γ is analytic, there is an $r<1$ such that φ can be extended to a conformal mapping of the unbounded component of the complement of a curve $\gamma_r\subseteq K^0:=\operatorname{int}(K)$ onto the exterior of the circle C_r (cf. [2, p. 45]). Clearly, if for $r\leqq\rho\leqq 1$, K_ρ denotes the compact set bounded by the curve $\gamma_\rho=\varphi^{-1}(C_\rho)$, then the Faber polynomials of $K_1=K$ and K_ρ are identical (in what follows we may assume $r<1$ so large that γ_ρ is a simple closed analytic curve for $r\leqq\rho\leqq 1$). But then there exist constants $A>0$ and $0<a<1$ such that on γ_1 the modulus of the difference between $F_n(z)$ and $\varphi^n(z)$ is at most Aa^n for all n (see [2, p. 108]).

We will show that for $\max\{a^{1/2},r\}<\rho<1$ all the zeros of $T^K_n=T_n$ lie in K_ρ for large n, and proving this will complete the sufficiency part. Let $a^{1/2}<b<\rho$. First we claim that for $z\in\gamma_1$ we have $|T_n(z)-F_n(z)|\leqq Bb^n$ for some constant B independent of z and n. To prove this claim we expand T_n in its Faber series:

$$T_n(z)=F_n(z)+c_1F_{n-1}(z)+\cdots+c_nF_0(z).$$

It is known (see [3, p. 58]) that the Fourier expansion of $T_n(\varphi^{-1}(e^{i\theta}))$ has the form

$$T_n(\varphi^{-1}(e^{i\theta}))\sim e^{ik\theta}+c_1e^{i(k-1)\theta}+\cdots+c_n+q_1e^{-i\theta}+\cdots$$

and so from the Parseval formula we get

$$1+|c_1|^2+\cdots+|c_n|^2\leqq\frac{1}{2\pi}\int_0^{2\pi}|T_n(\varphi^{-1}(e^{i\theta}))|^2\,d\theta.$$

We have already remarked that $|F_n(z)-\varphi^n(z)|\leqq Aa^n$ for $z\in\gamma_1=\partial K$, which implies that $\|F_n\|_K\leqq 1+Aa^n$, and so $\|T_n\|_K\leqq 1+Aa^n$. Substituting this into the previous estimate we get

$$|c_1|^2+\cdots+|c_n|^2\leqq 2Aa^n+A^2a^{2n},$$

from which the inequality $|T_n(z)-F_n(z)|\leqq D\sqrt{n}\cdot a^{n/2}$ immediately follows for $z\in\gamma_1$ with a constant D (note that the Faber polynomials are uniformly bounded on γ_1), and this proves our claim.

Next we note that the Bernstein-Walsh lemma (cf. [4, p. 77]) yields the following inequality for the supremum norms:

$$\|T_n - F_n\|_{\gamma_s} \leqq s^n \|T_n - F_n\|_{\gamma_1} \text{ for any } s \geqq 1,$$

and since on γ_s we have already seen that $|F_n| = s^n(1+o(1))$ uniformly in $s \geqq 1$, the inequality $|T_n(z) - F_n(z)| < |F_n(z)|$ follows for every large n, say $n \geqq n_0$, and any $z \notin K$. Hence T_n has no zeros outside K for large n.

Now let $b < b_1 < \rho$. From what we have discussed above concerning F_n and φ^n it also follows that for $z \notin K_\rho$ we have uniformly $|F_n(z)| \geqq db_1^n$ for some positive constant d, and at the same time $|T_n(z) - F_n(z)| \leqq Bb^n$ inside γ_1. Thus we can conclude again that T_n has no zero in $K \setminus K_\rho$ for large n.

This completes the sufficiency part of the proof.

(Necessity.) Now suppose that the zeros stay away from the boundary. Let

$$\nu_n := \frac{1}{n} \sum_{k=1}^{n} \delta_{z_k^{(n)}}$$

be the normalized counting measure on the zeros $z_k^{(n)}$ of T_n. Since the zeros of the T_n's lie in the convex hull of K, we can select a subsequence $\{\nu_{n_k}\}$ converging in the weak-star topology (on Borel measures with compact support) to some measure ν. According to our assumption and the Widom theorem mentioned in the Introduction, ν is supported in a compact subset of the interior K^0 of K. By assumption, K has connected interior and so there is a compact set $H \subset K^0$ such that H has connected interior containing the support of ν and $T_n(z) \neq 0$ for $z \in K \setminus H$ and n large. Let

$$g(z) := \exp\left(-\int \log \frac{1}{z-t}\, d\nu(t)\right),$$

where we take that branch of the logarithm that is positive for positive z. Then g is defined, analytic and single-valued in $\mathbb{C} \setminus H$ (note that ν is a probability measure). In $K \setminus H$ and also in a neighborhood of the boundary of K

$$|T_{n_k}(z)|^{1/n_k} \to \exp\left(\int \log|z-t|\, d\nu(t)\right), \tag{1}$$

and this combined with the fact that

$$\lim_{n\to\infty} \|T_n\|_K^{1/n} = \operatorname{cap}(K) \tag{2}$$

yields the result that the function

$$\log|g(z)| = \int \log|z-t|\, d\nu(t)$$

is a harmonic function in $\mathbb{C} \setminus H$, is of the form $\log|z| + o(1)$ around the infinity, and is at most as large as $\log(\operatorname{cap}(K))$ on $K \setminus H$. If $\mathcal{G}(z)$ denotes the Green's function with pole at infinity for the complement of K, then we have again $\mathcal{G}(z) + \log(\operatorname{cap}(K)) = \log|z| + o(1)$ as $z \to \infty$, but $\mathcal{G}(z) + \log(\operatorname{cap}(K)) \geqq \log(\operatorname{cap}(K))$ in $\mathbb{C} \setminus K$. Therefore, from the maximum principle for harmonic functions, we get first that $\mathcal{G}(z) + \log(\operatorname{cap}(K)) \geqq \log|g(z)|$ in $\mathbb{C} \setminus K$ and then that these two functions actually coincide because their difference is zero at infinity. From this we get that $\log|g(z)| > \log(\operatorname{cap}(K))$ outside K. In the interior of $K \setminus H$ we obtain from (1) and (2) and the

maximum principle that $\log|g(z)| < \log(\text{cap}(K))$. These facts imply that on the boundary of K we must have $\log|g(z)| = \log(\text{cap}(K))$ and that at no other point of $\mathbb{C}\setminus H$ can we have equality. Thus,

$$\partial K = \{z \in \mathbb{C}\setminus H \,|\, |g(z)| = \text{cap}(K)\}$$

and from this we will deduce that ∂K is in fact an analytic curve.

Without loss of generality we may assume $\text{cap}(K) = 1$. First of all we show that ∂K is locally an analytic curve. Let z_0 be an arbitrary point on the boundary of K. If $g'(z_0) \neq 0$, then g has an analytic inverse g^{-1} in a neighborhood U of z_0, and in this neighborhood ∂K coincides with the image of a portion of the unit circle under the mapping g^{-1}. Hence, for some neighborhood $U_1 \subset U$ of z_0, the intersection $\partial K \cap U_1$ is the analytic image of an arc on the unit circle, and so it is analytic.

Now suppose that $g'(z_0) = \cdots = g^{(k-1)}(z_0) = 0$, $g^{(k)}(z_0) \neq 0$, with $k \geqq 2$. Then g can be represented in a neighborhood U of z_0 as $g(z) = c + (h(z))^k$, where $|c| = |g(z_0)| = 1$, h is analytic in U, and $h(z_0) = 0$ but $h'(z_0) \neq 0$. For some small $\delta > 0$ the set

$$\{w \,|\, |c + w^k| = 1, |w| \leqq \delta\}$$

is the union of k analytic arcs intersecting the x axis at zero with angle $(\pi/2 + \arg c)/k + j\pi/k$, $0 \leqq j < k$. According to what we have said above, this implies that, in some neighborhood $U_1 \subset U$ of z_0, the part of the boundary ∂K lying in U_1 is the union of k analytic arcs such that their tangent lines at their common point z_0 divide the plane into $2k$ congruent sectors. Let γ_δ be the inverse image of the circle $|w| = \delta$ under the mapping $w = h(z)$, $z \in U_1$. Then it follows from h being conformal around z_0 that, for small $\delta > 0$, γ_δ is a simple closed curve such that ∂K divides it into $2k$ connected pieces: $\gamma_{\delta,0}, \cdots, \gamma_{\delta,2k-1}$, where each of these Jordan arcs is considered without its endpoints. Let $P_j \in \gamma_{\delta,j}$, $j = 0, \cdots, 2k-1$. Then P_0 belongs either to K^0 or to G_∞; for definiteness, suppose that $P_0 \in G_\infty$. As we move away from P_0 we stay in G_∞ until we reach ∂K. This implies that $P_1 \in K^0$, since in the opposite case we would have $P_1 \in G_\infty$, which would mean that the common endpoint S of $\gamma_{\delta,0}$ and $\gamma_{\delta,1}$ had a neighborhood disjoint from K^0, contradicting the maximum principle (recall that outside K^0 we have $|g| \geqq 1$ and that $|g(S)| = 1$ because $S \in \partial K$). In a similar fashion we can see that $P_2 \in G_\infty$ and $P_3 \in K^0$. Now, since G_∞ is connected, the points P_0 and P_2 can be joined by an arc Γ_∞ lying in G_∞. Since it is not possible to join P_0 and P_2 inside γ_δ (the possibility of joining P_1 and P_3 to z_0 in K^0 inside γ_δ prevents this), we can assume that Γ_∞ lies exterior to γ_δ (except for its endpoints). Similarly, since K^0 is connected, the points P_1 and P_3 can be joined by an arc Γ_0 in K^0 that also lies exterior to γ_δ. But clearly such a pair of arcs must intersect, which is absurd because $G_\infty \cap K^0 \neq \varnothing$. This contradiction shows that $g'(z_0) = 0$ cannot occur.

We have thus shown that ∂K locally is an analytic and simple curve. To complete the proof we have only to mention that ∂K must be connected because K^0 is connected.

REFERENCES

[1] H.-P. BLATT, E. B. SAFF, AND M. SIMKANI, *Jentzsch-Szegö type theorems for the zeros of best approximants*, J. London Math. Soc., 38 (1988), pp. 307-316.

[2] A. I. MARKUSHEVICH, *Theory of Functions of a Complex Variable*, Vol. III, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[3] CH. POMMERENKE, *Univalent Functions*, Vandenhoeck and Ruprecht, Göttingen, F.R.G., 1975.

[4] J. L. WALSH, *Interpolation and Approximation by Rational Functions in the Complex Domain*, Amer. Math. Soc. Colloq. Publ., American Mathematical Society, Providence, RI, 1935 (Fifth edition, 1969).

[5] H. WIDOM, *Polynomials associated with measures in the complex plane*, J. Math. Mech., 16 (1967), pp. 997-1013.

SIAM J. MATH. ANAL.
Vol. 21, No. 3, pp. 803–821, May 1990
© 1990 Society for Industrial and Applied Mathematics
016

A PROOF OF THE MACDONALD–MORRIS ROOT SYSTEM CONJECTURE FOR F_4*

F. G. GARVAN†

Abstract. This paper gives a proof of the $q=1$ case of the Macdonald–Morris root system conjecture for F_4 that draws on ideas from Zeilberger's recent proof of the $G_2^\vee$ case and Kadell's proof of the q-BC_n case. The present proof depends on much computer computation. As in Zeilberger's proof, the problem is reduced to solving a system of linear equations. A FORTRAN program generated the equations, which were solved using the computer algebra package MAPLE.

Key words. beta integrals, constant terms, computer algebra, Macdonald's conjectures, Macdonald–Morris conjectures, Mehta's integral, multidimensional integrals, root systems, Selberg's integral

AMS(MOS) subject classifications. primary 33A15; secondary 33A75, 17B20

1. Introduction.

In 1962, Dyson [D] in some statistical work in nuclear physics conjectured

$$(D') \qquad (2\pi)^{-n}\int_0^{2\pi}\cdots\int_0^{2\pi}\prod_{1\le i<j\le n}\left|e^{i\theta_i}-e^{i\theta_j}\right|^{\beta}d\theta_1\cdots d\theta_n=\frac{\Gamma(1+\frac{1}{2}n\beta)}{\Gamma(1+\frac{1}{2}\beta)^n}$$

$$(\mathrm{Re}\,(\beta)>0).$$

In the electrostatic analogue the left side is the positional partition function of n point charges in a unit circle with angular variables $[\theta_1,\cdots,\theta_n]$ and $\beta=1/T$, where T is the temperature at thermal equilibrium. By appealing to Carlson's theorem [T, p. 186] we can assume without loss of generality that $\beta=2k$ is an even nonnegative integer. Now, by the orthogonality of the exponentials, Dyson observed that (D′) can be written as the constant term of a Laurent polynomial:

$$(D) \qquad \text{Constant term of}\prod_{1\le i\ne j\le n}\left(1-\frac{x_i}{x_j}\right)^k=\frac{(kn)!}{(k!)^n}.$$

Since each term in the product on the left can be expanded by the binomial theorem we can view (D) as a multisum binomial coefficient identity, or equivalently, as a terminating hypergeometric multiple series identity. Such series rarely factor and when they do, as in (D), it is interesting and important. The $n=2$ case follows trivially. Dyson proved the $n=3$ case from Dixon's [D, p. 152] identity for the alternating sum of the cubes of binomial coefficients. This identity is a special case of Dixon's [Ba, § 3.1] evaluation of a well-poised ${}_3F_2$. This led Dyson to add more parameters to his conjecture:

$$(\text{general-D}) \qquad \text{C.T.}\prod_{1\le i\ne j\le n}\left(1-\frac{x_i}{x_j}\right)^{a_i}=\frac{(a_1+\cdots+a_n)!}{a_1!\cdots a_n!}$$

where a_i are nonnegative integers. Here we use C.T. as our abbreviation for *constant term.* This more general conjecture was proved in [Gu], [W], and [Go]. A q-analogue of (general-D) was conjectured by Andrews [An] and later proved by Zeilberger and Bressoud [Z-B].

* Received by the editors July 19, 1988; accepted for publication (in revised form) March 6, 1989.

† School of Mathematics and Physics, Macquarie University, New South Wales 2109, Australia. This research was done at the University of Wisconsin, Madison, and later while the author was a postdoctoral member of the Institute of Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota.

If the unit circle in Dyson's problem above is replaced by the real line then the analogue of (D') is the Mehta-Dyson conjecture [Me], [M-D]

$$\text{(M-D)} \qquad (2\pi)^{-n/2}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{i=1}^{n}e^{-t_i^2/2}\prod_{1\leqq i<j\leqq n}|t_i-t_j|^{2z}\,dt_1\cdots dt_n=\prod_{j=1}^{n}\frac{\Gamma(jz+1)}{\Gamma(z+1)}.$$

This was proved by taking the appropriate limits in Selberg's [Se] integral:

$$\text{(S)} \qquad \begin{aligned}&\int_0^1\cdots\int_0^1\prod_{i=1}^{n}t_i^{x-1}(1-t_i)^{y-1}\prod_{1\leqq i<j\leqq n}|t_i-t_j|^{2z}\,dt_1\cdots dt_n\\&\quad=\prod_{j=1}^{n}\frac{\Gamma(x+(j-1)z)\Gamma(y+(j-1)z)\Gamma(jz+1)}{\Gamma(x+y+(n+j-2)z)\Gamma(z+1)}\end{aligned}$$

(see [Ma2]).

In 1982, Macdonald [Ma2] conjectured generalizations of both (D) and (M-D) in the context of root systems of Lie algebras. The generalization of (D) is given below in (M). The BC_n case is intimately connected to Selberg's integral (see [Ma2] and [As]). Recently much progress has been made on q-generalizing and extending Selberg's integral. Aomoto [Ao] has given a new proof that depends on little more than integration by parts; see [As] for a nice account. Kadell [K1] and Habsieger [Hab1] have proved Askey's q-analogue of Selberg's integral. Homogeneous symmetric polynomials $Z_\mu(t; z)$ have been added to the integrand of Selberg's integral so that the resulting integral still has closed form (see [Ri] and for the q case [K3]). For $z=\frac{1}{2}$ they are the zonal polynomials of statistics studied by James ([Jam1], [Jam2], [Jam3]) and Constantine [Co]. For $z=1$ they are Schur functions and for arbitrary z they are the Jack polynomials [Jac] studied by Macdonald [Ma3], Stanley [St], and Kadell [K3]. It is hoped that some of these results may be extended to other root systems yielding further applications to multivariate statistics. See [He], [O1] for a setting for generalized hypergeometric functions associated with root systems.

Recently the Macdonald conjectures have led to interesting developments in other areas of mathematics. In combinatorics, a study of the A_n and q-A_n cases in [Z1], [Z-B] led to the technique of counting and q-counting tournaments. In Lie algebra, Hanlon [Han1], [Han2] has found an interesting formulation and refinement of the Macdonald conjectures in the context of the cyclic homology of the exterior product of a Lie algebra with $\mathbb{C}[t, t^{-1}]$. In number theory, character sum analogues of the Macdonald conjectures have been found by Evans [E]. In algebra, Regev [Re1], [Re2] has found connections between Mehta-type integrals, PI rings, and representations of the symmetric group.

We now describe the simplest form of Macdonald's generalization of Dyson's conjecture (D). Unfortunately, no generalization to all root systems is known for the more general (general-D). Let R be a reduced root system, let x^α denote the formal exponential corresponding to $\alpha\in R$, and let k be a nonnegative integer. Then Macdonald conjectured that

$$\text{(M)} \qquad \text{C.T.}\prod_{\alpha\in R}(1-x^\alpha)^k=\binom{kd_1}{k}\cdots\binom{kd_l}{k}.$$

Here C.T. means constant term in the Laurent polynomial in the $x^{\pm\alpha}$ and the d_i are the degrees of the fundamental invariants of the Weyl group of R. Dyson's conjecture is equivalent to the A_{n-1} case. Morris [Mo] conjectured a generalization of (M) for the G_2 case with an extra parameter. This led to the Macdonald-Morris conjecture ([Ma2], p. 988), which has the same form as (M) except that k is replaced by k_α (with

the restriction that k_α is constant on roots of equal length) and the binomial coefficients on the right side are replaced by certain factorials. Also in [Ma2] q-analogues are given. These q-Macdonald-Morris conjectures are cast in the language of affine root systems [Ma1] and are related to the Macdonald identities of [Ma1]. In this paper we restrict attention to the $q=1$ case.

The Macdonald-Morris conjectures are known for A_n ([Gu], [W], [Go]), B_n, C_n, D_n, BC_n (follow from Selberg's [Se] integral as noted by Macdonald [Ma2]), G_2 ([Z2], [Hab2]). The goal of this paper is to give a proof of the F_4 case spelled out below in (1.1). Since this paper was first written the Macdonald-Morris conjectures have been proved for all root systems by Opdam [O2]. We should mention that recently the $G_2^\vee$ ([Z3]) and the BC_n ([K2]) cases of q-Macdonald-Morris have been proved.

Opdam's proof involves using his shift operators [O1] for general root systems and Heckman's [He] L_2-norm formula for the orthogonal polynomials associated with root systems. Macdonald [Ma2] has also generalized the Mehta-Dyson conjecture (M-D) to finite Coxeter groups. Opdam [O2] has proved the Weyl group case. Also, since this paper was first written, we have been able to extend the methods of this paper to prove the $S(F_4)$ and $S(F_4)^\vee$ cases of the q-Macdonald-Morris conjectures. (This should appear in a joint paper with G. Gonnet.) We have also been able [Ga2] to prove the icosahedral case I_3 of the Macdonald-Mehta-Dyson integral which is not accessible via Opdam's method. Moreover, we have been able to show [Ga3] that all cases of the q-Macdonald-Morris conjecture and the Macdonald-Mehta-Dyson integrals can be written in *closed form.* Although Opdam's proof is very beautiful it does not include all the results of this paper (see, for instance, § 8). We also hope that a more elementary proof is possible.

The Macdonald-Morris conjecture for F_4 is

$$\begin{aligned}
&\text{C.T.} \prod_{1\le i<j\le 4}(1-x_ix_j)^a(1-x_i^{-1}x_j^{-1})^a\left(1-\frac{x_i}{x_j}\right)^a\left(1-\frac{x_j}{x_i}\right)^a\\
&\quad\cdot\prod_{i=1}^4(1-x_i^2)^b(1-x_i^{-2})^b\prod_{r_1,r_2,r_3,r_4=\pm 1}(1-x_1^{r_1}x_2^{r_2}x_3^{r_3}x_4^{r_4})^b\\
&=\frac{(6a+6b)!(4a+4b)!(2a+6b)!(4a+2b)!(2a+4b)!(4b)!(3a)!}{(5a+6b)!(3a+5b)!(3a+4b)!(3a+2b)!(2a+3b)!(a+3b)!(2a+b)!}\\
&\quad\cdot\frac{(3b)!(2a)!(2b)!}{(a+2b)!(a+b)!a!a!b!b!b!}=f(a,b).
\end{aligned}\tag{1.1}$$

The goal of this paper is to prove (1.1). Kadell's paper [K2, § 2] also contains a new proof of the $(q=1)$ BC_n case of the Macdonald-Morris conjecture that avoids integrals. This new proof is analogous to Aomoto's [Ao] proof of Selberg's integral in the following sense: It involves adding extra factors to the Laurent polynomial as opposed to adding extra factors to the integrand of Selberg's integral, and Aomoto's integration by parts is replaced by the fact that the derivative of a Laurent polynomial has no residue. We extend Kadell's proof to the F_4 case, but another idea is needed. The extra idea comes from Zeilberger's [Z3] proof of the $G_2^\vee$ case of the q-Macdonald-Morris conjecture. In [Z3] Zeilberger describes a method for handling q-Macdonald-Morris for a specific root system given enough computer time, memory, and luck. As in [Z3] the problem is reduced to finding and solving a system of linear equations whose unknowns are constant terms of certain Laurent polynomials. These equations are generated with the aid of a FORTRAN program. Finally the equations are solved using the computer algebra package MAPLE.

After some preliminaries in § 2 an idea of the proof is given in § 3. The results behind the FORTRAN program that generates the desired equations are given in §§ 4–6. The proof is completed in § 7.

We have been able to verify the results of this paper by another method. Recently [Ga1] we have found a new proof of the G_2 case of the Macdonald–Morris conjecture which is solely in terms of integrals. Our proof was motivated by some conjectures of Askey [As] that have to do with adding roots to the G_2 case of the Macdonald–Morris conjecture, and is analogous to Aomoto's [Ao] proof of Selberg's integral. We have been able to extend our integral proof of the G_2 case to the F_4 case. However, this other proof involves finding equations between certain integrals and converting these into equations involving constant terms. The proof then proceeds as usual by solving a system of linear equations. We have omitted this other proof, finding the approach of working with Laurent polynomials rather than with integrals more straightforward.

In § 8 we give some other results that involve adding extra factors to the F_4 case of the Macdonald–Morris conjecture. The results are analogous to Askey's [As] conjectures for G_2. Although many of these results can be written as products of factorials, we are unable to generalize them to all root systems. These other results may indicate that a simpler proof of the F_4 case is possible.

All computer programs used in this paper are available from the author on request. Some preliminary calculations were done using REDUCE at the University of Wisconsin, Madison. The final FORTRAN and MAPLE programs were run on an APOLLO DN-5800 at the Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis.

2. Some preliminaries. In this section we prove some properties of the root system F_4 that will be needed later. We assume that the reader is familiar with the basics of root systems and their Weyl groups. See [Bo], [Ca], and [Hu] for treatments of root systems and Weyl groups.

Let $\{e_1, e_2, e_3, e_4\}$ be the standard basis of $\mathbb{R}^4$. The roots of F_4 are usually written as

$$\pm e_i \quad (1 \leqq i \leqq 4), \quad \pm e_i \pm e_j \quad (1 \leqq i < j \leqq 4), \quad \tfrac{1}{2}(\pm e_1 \pm e_2 \pm e_3 \pm e_4); \tag{2.1}$$

see, for example, [Bo, p. 272]. We call this set of roots $\Phi^{(1)}(F_4)$. It is clear that the long roots of F_4 are isomorphic to D_4. In this paper we shall use two other ways of writing the roots of F_4.

First, we rewrite the roots of F_4 to make it clear that the short roots of F_4 are isomorphic to D_4. Let $\Phi^{(2)}(F_4)$ be the set of vectors

$$\pm 2e_i \quad (1 \leqq i \leqq 4), \quad \pm e_i \pm e_j \quad (1 \leqq i < j \leqq 4), \quad (\pm e_1 \pm e_2 \pm e_3 \pm e_4). \tag{2.2}$$

$\Phi^{(2)}(F_4)$ and $\Phi^{(1)}(F_4)$ are isomorphic as root systems. The isomorphism is given by

$$A: \Phi^{(1)}(F_4) \to \Phi^{(2)}(F_4) \tag{2.3}$$

where A is the transformation with matrix

$$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \tag{2.4}$$

with respect to the standard basis of $\mathbb{R}^4$. This is a root system isomorphism since $A^tA = 2I$. As an immediate consequence we have the following lemma.

LEMMA 2.5. *Short roots of* $F_4 \cong D_4$.

Second, we may write the roots of F_4 (as given in (2.2)) as $\mathbb{Z}$-linear combinations of $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, where

$$\alpha_1 = e_1 - e_2, \quad \alpha_2 = e_3 + e_4, \quad \alpha_3 = e_3 - e_4, \quad \alpha_4 = e_2 - e_3. \tag{2.6}$$

These $\mathbb{Z}$-linear combination are given in Appendix A. The α_i come from the Dynkin diagram for D_4, which is given in Fig. 1.

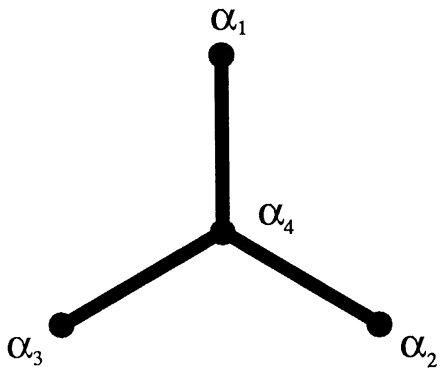


FIG. 1. *Dynkin diagram for* D_4.

From the symmetry of the Dynkin diagram we see that any permutation of $\alpha_1, \alpha_2, \alpha_3$ leaves the root system of D_4 invariant. It is interesting to note that any such permutation also leaves the root system of F_4 invariant. Let $\pi \in S_3$ and suppose $\alpha = \sum_{i=1}^4 k_i\alpha_i \in \Phi^{(2)}(F_4)$, where $k_i \in \mathbb{Z} (1 \leqq i \leqq 4)$. Then we define

$$\pi\alpha = k_1\alpha_{\pi(1)} + k_2\alpha_{\pi(2)} + k_3\alpha_{\pi(3)} + k_4\alpha_4. \tag{2.7}$$

We have the following lemma.

LEMMA 2.8. *For* $\pi \in S_3$,

$$\pi(\Phi^{(2)}(F_4)) = \Phi^{(2)}(F_4).$$

For a root system R we denote its Weyl group by $W(R)$. We need a nice way to code the elements of $W(F_4)$. By [Bo, p. 257], $W(D_4)$ consists of all signed permutations, with an even number of signs, that act on the coordinates e_1, e_2, e_3, e_4. Let H denote the set of all signed permutations that act on the coordinates e_1, e_2, e_3, e_4. For $\alpha \in R$ we denote by w_α the reflection through the hyperplane orthogonal to α. Since $w_{2e_i} \in W(F_4) (1 \leqq i \leqq 4)$ we have

$$W(D_4) \subset H \subset W(F_4). \tag{2.9}$$

We introduce some notation to describe the elements of H. We denote the permutations on the coordinates e_1, e_2, e_3, e_4 using the usual cycle notation. We define the sign changes as follows: For $1 \leqq i \leqq 4$ let s_i denote the transformation given by

$$s_i : \mathbb{R}^4 \longrightarrow \mathbb{R}^4, \quad e_i \mapsto -e_i, \quad \text{and} \quad e_j \mapsto e_j \quad (j \neq i). \tag{2.10}$$

For example,

$$s_1(34)(e_1 + e_2 + 2e_3 - e_4) = -e_1 + e_2 - e_3 + 2e_4.$$

LEMMA 2.11. *Every* $w \in W(F_4)$ *can be written*

$$w = (\tau\sigma)^k h$$

where $k = 0, 1, 2$, $h \in H$, $\tau = w_{2e_4}$ *and* $\sigma = w_{e_1-e_2-e_3-e_4}$.

Proof. We have

$$|H| = 2^4.4! = 2^7.3, \qquad |W(F_4)| = 2^7.3^2$$

[Bo, p. 273, Eq. (X)].

The result follows since $\tau\sigma \notin H$ and $(\tau\sigma)^3 = I$. □

Let

$$\bar{C} = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 : x_1 \geqq x_2 \geqq x_3 \geqq x_4 \geqq 0, \quad x_2 + x_3 + x_4 \geqq x_1, x_1 + x_4 \geqq x_2 + x_3\}. \tag{2.12}$$

LEMMA 2.13. *For every* $x \in \mathbb{R}^4$ *there is a unique* $v \in \bar{C}$ *such that* x *can be transformed into* v *by some element of* $W(F_4)$.

Proof. From [Bo, p. 272, Eq. (II)] the following vectors form a base for F_4:

$$e_2 - e_3, \quad e_3 - e_4, \quad e_4, \quad \tfrac{1}{2}(e_1 - e_2 - e_3 - e_4), \tag{2.14}$$

with the roots written in the usual way (i.e., as elements of $\Phi^{(1)}(F_4)$). In our representation of F_4 (i.e., as elements of $\Phi^{(2)}(F_4)$) this corresponds to

$$e_1 - e_2 + e_3 - e_4, \quad -e_1 + e_2 - e_3 - e_4, \quad e_1 + e_4, \quad -e_1 + e_3. \tag{2.15}$$

By applying the relevant signed permutation we find that the following vectors also form a base for F_4:

$$-e_1 + e_2 + e_3 + e_4, \quad e_1 - e_2 - e_3 + e_4, \quad e_3 - e_4, \quad -e_3 + e_2, \tag{2.16}$$

by [Ca, Thm. 2.2.4]. The closure of the chamber C corresponding to this base is given in (2.12). The result follows from [Ca, Prop. 2.3.4]. □

Let

$$\text{funch}: \mathbb{R}^4 \longrightarrow \bar{C}, \qquad \text{funch}(x) = v \tag{2.17}$$

with x, v as in Lemma 2.13 above. For $\alpha = \sum_{i=1}^4 k_i e_i (k_i \in \mathbb{Z})$ we let

$$x^\alpha = \prod_{i=1}^4 x_i^{k_i}. \tag{2.18}$$

The elements w of the Weyl group act on monomials by

$$w(x^\alpha) = x^{w(\alpha)}, \tag{2.19}$$

and by linearity on Laurent polynomials that are linear combinations of the x^α.

Let

$$F(x; a, b) = \prod_{\alpha \in \Phi^{(2)}(F_4)} (1 - x^\alpha)^{k_\alpha} \tag{2.20}$$

where

$$k_\alpha = \begin{cases} a & \text{if } \alpha \text{ is a short root,} \\ b & \text{if } \alpha \text{ is a long root.} \end{cases}$$

We note that $F(x; a, b)$ is the Laurent polynomial on the left-hand side of (1.1).

LEMMA 2.21. *For* $w \in W(F_4)$,

$$\text{C.T. } x^{\alpha} F = \text{C.T. } x^{w(\alpha)} F$$

where F *is defined in* (2.20).

Proof. The result follows from the fact that F is symmetric with respect to the Weyl group and w does not change the constant term. □

3. The idea of the proof. Let

$$f'(a, b) = \text{C.T. } F(x; a, b) \tag{3.1}$$

where F is defined in (2.20). Our goal is to prove that $f'(a, b) = f(a, b)$ for all $a, b \geqq 0$. The idea is to proceed by induction on a. That is, we want to prove that

$$\frac{f'(a+1, b)}{f'(a, b)} = \frac{f(a+1, b)}{f(a, b)}. \tag{3.2}$$

This will be enough because the case $a = 0$ is already known, since

$$\text{Long roots of } F_4 \cong D_4. \tag{3.3}$$

The flavor of our proof is similar to Zeilberger's [Z3] proof of the $G_2^{\vee}$ case of the q-version of the Macdonald-Morris root system conjecture. Let L be the lattice generated by $\alpha (\alpha \in F_4)$. Now,

$$\begin{aligned}
f'(a+1, b) &= \text{C.T. } F(x; a+1, b) \\
&= \text{C.T. } \prod_{\alpha \in \text{short } F_4} (1 - x^{\alpha}) F(x; a, b) \\
&= \text{C.T. } \prod_{\alpha \in D_4} (1 - x^{\alpha}) F(x; a, b) \qquad \text{(by Lemma 2.5)} \\
&= \text{C.T. } \sum_{\alpha \in L'} a_{\alpha} x^{\alpha} F(x; a, b) \qquad \text{(for some } L' \subset L\text{)} \\
&= \text{C.T. } \sum_{\alpha \in L'} a_{\alpha} x^{\text{funch}(\alpha)} F(x; a, b) \qquad \text{(by (2.17) and Lemmas 2.13, 2.21)} \\
&= \text{C.T. } \sum_{\alpha \in S} a'_{\alpha} x^{\alpha} F(x; a, b)
\end{aligned} \tag{3.4}$$

for some finite subset S of $\bar{C}$, which is defined in (2.12). $\prod_{\alpha \in D_4} (1 - x^{\alpha})$ was multiplied out using a FORTRAN program; funch (α) was calculated for each monomial x^{α} that arose in each stage of the multiplication. The 37 vectors that arose, $v(i) (1 \leqq i \leqq 37)$, are listed in Appendix B. $S = \{v(i): 1 \leqq i \leqq 37\}$ and

$$\sum_{\alpha \in S} a'_{\alpha} x^{\alpha} = 192 x^{v(1)} - 768 x^{v(2)} + \cdots + 192 x^{v(37)}. \tag{3.5}$$

The complete list of coefficients in (3.5) is given in Appendix B. Let

$$an(i) = \text{C.T. } x^{v(i)} F(x; a, b) \quad (1 \leqq i \leqq 37). \tag{3.6}$$

The problem is to get $an(i) (2 \leqq i \leqq 37)$ in terms of $an(1) = \text{C.T. } x^{v(1)} F(x; a, b) = \text{C.T. } F(x; a, b) = f'(a, b)$. Once we have done this, (3.2) should follow from (3.4) and (3.5).

Hence we need to find 36 independent equations in the unknowns $an(i) (1 \leqq i \leqq 37)$. Our goal is to write a FORTRAN program that will generate equations. The input of this program is a 4-tuple $k = (k_1, k_2, k_3, k_4) \in \mathbb{N}^4$. The output will be either a homogeneous linear equation in the $an(i) (1 \leqq i \leqq 37)$ or an error message, which says such an equation is not possible. It will turn out that, when $k = v(i) (2 \leqq i \leqq 37)$, the

corresponding outputs will be the required 36 independent equations in the $an(i)(1 \leqq i \leqq 37)$. See § 6 for more details. We describe how we came by this program in four steps:
Step 1. Use

$$\text{C.T. } x_1 \frac{\partial}{\partial x_1} x_1^{k_1} x_2^{k_2} x_3^{k_3} x_4^{k_4} F(x; a, b) = 0. \tag{3.7}$$

As we noted before, this idea was used by Kadell [K2] in his proof of the q-BC_n case. This gives rise to an equation involving constant terms of rational functions in x_1, x_2, x_3, x_4 times F.

Step 2. Use the Weyl group to reduce the number of types of terms arising in Step 1.

Step 3. Use the Weyl group to write the constant terms that arise in Step 2 as constant terms of Laurent polynomials times F.

Step 4. Use the Weyl group to write the constant terms that arise in Step 3 in terms of the $an(i)(1 \leqq i \leqq 37)$, if possible.

In this way, for certain k, (3.7) can be written as a linear homogeneous equation in the $an(i)(1 \leqq i \leqq 37)$.

4. Steps 1 and 2. An equation involving constant terms with denominators. In this section we describe the first two steps, mentioned in § 3, that are needed in turning (3.7) into an equation involving the $an(i)$. Equation (3.7) is

$$\begin{aligned} 0 = \text{C.T.} \Bigg\{ \Bigg(k_1 + a \sum_{j=2}^{4} \Bigg(& \frac{-x_1 x_j}{1-x_1 x_j} + \frac{x_1^{-1} x_j^{-1}}{1-x_1^{-1} x_j^{-1}} - \frac{x_1/x_j}{1-x_1/x_j} + \frac{x_j/x_1}{1-x_j/x_1} \Bigg) \\ & + b \Bigg(\frac{-2x_1^2}{1-x_1^2} + \frac{2x_1^{-2}}{1-x_1^{-2}} \Bigg) \\ & + b \sum_{r_2, r_3, r_4 = \pm 1} \Bigg(\frac{-x_1 x_2^{r_2} x_3^{r_3} x_4^{r_4}}{1-x_1 x_2^{r_2} x_3^{r_3} x_4^{r_4}} + \frac{x_1^{-1} x_2^{r_2} x_3^{r_3} x_4^{r_4}}{1-x_1^{-1} x_2^{r_2} x_3^{r_3} x_4^{r_4}} \Bigg) \Bigg) \cdot x^k F(x; a, b) \Bigg\} \end{aligned} \tag{4.1}$$

where $k = (k_1, k_2, k_3, k_4)$. Hence,

$$\begin{aligned} \text{C.T. } k_1 x^k F = \text{C.T.} \Bigg(& a \sum_{j=2}^{4} \Bigg(\frac{1+x_1 x_j}{1-x_1 x_j} + \frac{1+(x_1/x_j)}{1-(x_1/x_j)} \Bigg) \\ & + 2b \frac{(1+x_1^2)}{(1-x_1^2)} + b \sum_{r_2, r_3, r_4 = \pm 1} \frac{1+x_1 x_2^{r_2} x_3^{r_3} x_4^{r_4}}{1-x_1 x_2^{r_2} x_3^{r_3} x_4^{r_4}} \Bigg) \cdot x^k F. \end{aligned} \tag{4.2}$$

This completes Step 1.

In Step 2 we reduce the number of different denominators appearing in Step 1 to two (one for each root length). The reason this can be done is that each term on the right-hand side of (4.2) is of the form $1+x^\alpha/1-x^\alpha$ for some $\alpha \in F_4$, and so can be converted to one of two types by using the fact that the Weyl group acts transitively on roots of equal length and by Lemma 2.21. Hence for each $\alpha \in F_4$ we need to find a $w \in W(F_4)$ such that

$$w(\alpha) = \begin{cases} e_1 - e_3, & \alpha \text{ short}, \\ e_1 - e_2 + e_3 + e_4, & \alpha \text{ long}. \end{cases}$$

It is clear that for α of the form $e_i \pm e_j$, $e_1 \pm e_2 \pm e_3 \pm e_4$, we may take $w \in H$, the set of signed permutations (defined in § 2), and w is easy to calculate. All that remains is to find a $w \in H$ such that $w(2e_1) = e_1 - e_2 + e_3 + e_4$. Let SYM denote the reflection through the hyperplane orthogonal to $e_1 - e_2 - e_3 - e_4$; then

$$2e_1 \xrightarrow{(13)} 2e_3 \xrightarrow{\text{SYM}} e_1 - e_2 + e_3 - e_4 \xrightarrow{s_4} e_1 - e_2 + e_3 + e_4.$$

Hence we find that (4.2) may be written as

$$
\begin{aligned}
0 = \text{C.T.}\Big\{ & k_1 x^k + a\frac{(x_1+x_3)}{(x_1-x_3)}(x^k+(23)x^k+(34)x^k \\
& + s_3x^k+(23)s_2x^k+(34)s_4x^k) \\
& + b\frac{(x_2+x_1x_3x_4)}{(x_1x_3x_4-x_2)}(2s_4\,\text{SYM}\,(13)x^k \\
& + s_2x^k+s_4s_2x^k+s_2s_3x^k+s_2s_3s_4x^k+x^k+s_4x^k \\
& + s_3x^k+s_3s_4x^k)\Big\} F(x;\, a,\, b).
\end{aligned} \tag{4.3}
$$

5. Step 3. Getting rid of denominators. The constant term expressions that arise in Step 2 can be written as either

$$\text{C.T.}\frac{p_1(x)}{x_1-x_3}F(x;\, a,\, b) \tag{5.1}$$

or

$$\text{C.T.}\frac{p_2(x)}{x_1x_3x_4-x_2}F(x;\, a,\, b) \tag{5.2}$$

where $p_i(x)(i=1,2)$ are Laurent polynomials. In this section we show how each of these expressions can be written in the form

$$\text{C.T.}\, p(x)F(x;\, a,\, b),$$

for some Laurent polynomial $p(x)$, and how such an expression can be computed.

Expression (5.1) is easy to handle. Since $F(x;\, a,\, b)$ is symmetric in x_1, x_3 we have

$$\text{C.T.}\frac{p_1(x)}{x_1-x_3}F(x;\, a,\, b) = \frac{1}{2}\text{C.T.}\left(\frac{p_1(x)-(13)p_1(x)}{x_1-x_3}\right)F(x;\, a,\, b) \tag{5.3}$$

and it is clear that $(p_1(x)-(13)p_1(x))/(x_1-x_3)$ is a polynomial.

Before we can handle (5.2) we need to define an algorithm, FUN, whose input is a vector given in terms of the $e_i(1 \leqq i \leqq 4)$ and whose output is the same vector given in terms of the $\alpha_i(1 \leqq i \leqq 4)$, defined in (2.6). FUN: $\mathbb{R}^4 \to \mathbb{R}^4$ is a linear transformation whose matrix is

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ 1 & 1 & 0 & 0 \end{pmatrix} \tag{5.4}$$

with respect to the bases $\{e_i\colon 1 \leqq i \leqq 4\}$, $\{\alpha_i\colon 1 \leqq i \leqq 4\}$. The inverse of FUN is UNFUN, with matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 1 & 1 & -1 \\ 0 & 1 & -1 & 0 \end{pmatrix}. \tag{5.5}$$

For $\alpha = \sum_{i=1}^{4} l_i \alpha_i (l_i \in \mathbb{Z})$ we let

$$y^{\alpha} = \prod_{i=1}^{4} y_i^{l_i}. \tag{5.6}$$

FUN acts on monomials by

$$\text{FUN}(x^{\alpha}) = y^{\text{FUN}(\alpha)}, \tag{5.7}$$

and by linearity on Laurent polynomials that are linear combinations of the x^{α}. For example,

$$\begin{aligned} \text{FUN}\left(\frac{x_1x_3x_4}{x_2}\right) &= \text{FUN}(x^{e_1-e_2+e_3+e_4}) \\ &= y^{\text{FUN}(e_1-e_2+e_3+e_4)} \\ &= y^{\alpha_1+\alpha_2} \\ &= y_1y_2. \end{aligned}$$

Alternatively, we could describe the action of FUN as replacing x_1 by $y_1y_4\sqrt{y_2y_3}$, x_2 by $y_4\sqrt{y_2y_3}$, x_3 by $\sqrt{y_2y_3}$, and x_4 by $\sqrt{y_2/y_3}$.

From Lemma 2.8 it follows that

$$F'(y; a, b) = \text{FUN}(F(x; a, b)) \tag{5.8}$$

is symmetric in y_1, y_2, y_3. Now we can handle (5.2):

$$\begin{aligned} &\text{C.T.} \frac{p_2(x)}{x_1x_3x_4 - x_2} F(x; a, b) \\ &= \text{C.T.} \frac{((12)p_2(x))}{x_2x_3x_4 - x_1} F(x; a, b) \quad \text{(by Lemma 2.21)} \\ &= \text{C.T.} \left(\frac{x_1^{-1}(12)p_2(x)}{x^{\alpha_2-\alpha_1}-1}\right) F(x; a, b) \\ &= \text{C.T.} \frac{q(y)}{y_2 - y_1} F'(y; a, b) \quad \text{(by applying FUN)} \end{aligned} \tag{5.9}$$

where $q(y)$ is the Laurent polynomial

$$q(y) = y_1 \,\text{FUN}(x_1^{-1}(12)p_2(x)). \tag{5.10}$$

Hence,

$$\begin{aligned} \text{C.T.} \frac{p_2(x)}{x_1x_3x_4 - x_2} F(x; a, b) &= \frac{1}{2} \text{C.T.} \left(\frac{q(y) - (12)q(y)}{y_2 - y_1}\right) F'(y; a, b) \\ &= \frac{1}{2} \text{C.T. UNFUN} \left(\frac{q(y) - (12)q(y)}{y_2 - y_1}\right) F(x; a, b) \end{aligned} \tag{5.11}$$

and UNFUN $((q(y) - (12)q(y))/(y_2 - y_1))$ is a Laurent polynomial in $x_i (1 \leqq i \leqq 4)$ as required.

We note that $(p_1(x) - (13)p_1(x))/(x_1 - x_3)$ and similarly UNFUN $((q(y) - (12)q(y))/(y_2 - y_1))$ can be easily computed by observing that if $p_1(x) = \sum_a c_a x^a = \sum_a c_a x_1^{a_1} x_2^{a_2} x_3^{a_3} x_4^{a_4}$ then

$$\frac{p_1(x) - (13)p_1(x)}{x_1 - x_3} = \sum_a \text{sgn}(a_1 - a_3) c_a x_2^{a_2} x_4^{a_4} \sum_{l=0}^{|a_1-a_3|-1} x_1^{\min(a_1,a_3)+l} x_3^{\max(a_1,a_3)-l-1}. \tag{5.12}$$

6. Step 4. Obtaining equations in the $an(i)$. In this section we describe the final step needed in converting the constant term equation (3.7) into an equation involving the $an(i)$. An examination of Steps 1–4 will then yield an algorithm whose input is a 4-tuple $k=(k_1, k_2, k_3, k_4)\in\mathbb{N}^4$ and whose output is an equation involving the $an(i)$, or an error message.

By Step 3 we can write the constant term equation (3.7), in terms of constant terms of certain Laurent polynomials times $F(x; a, b)$. In Step 4 we would like to find such expressions in terms of the $an(i)$. To do this we use funch, defined in (2.17). Suppose we are given such an expression, say $p(x)F(x; a, b)$, where

$$p(x)=\sum_{\alpha\in L'} a_\alpha x^\alpha \quad \text{(for some finite set } L'\text{)}. \tag{6.1}$$

Then

$$\begin{aligned}
&\text{C.T. } p(x)F(x; a, b)\\
&\quad=\text{C.T.}\sum_{\alpha\in L'} a_\alpha x^{\text{funch}(\alpha)}F(x; a, b) && \text{(by Lemma 2.13, (2.17), and Lemma 2.21)}\\
&\quad=\text{C.T.}\sum_{\alpha\in S} a'_\alpha x^\alpha F(x; a, b) && \text{(where } S\subseteq \text{funch}(L')\subseteq \bar{C} \text{ defined in (2.12))}\\
&\quad=\sum_{i=1}^{37} a'_{v(i)}an(i) && \text{(where } an(i) \text{ is defined in (3.6)).}
\end{aligned} \tag{6.2}$$

Note that here we are assuming $S\subseteq\{v(i)\}_{i=1}^{37}$, which may not necessarily be the case. However, for all values of the input k that we use, this condition is satisfied. We leave it to the reader to write a subroutine that will do the reduction described in (6.2). This subroutine should check whether $S\subseteq\{v(i)\}_{i=1}^{37}$. If this condition is not satisfied, the output of the subroutine should be some error message.

7. Step 5. Generating the equations and completing the proof. In the previous section we noted that Steps 1–4 yield an algorithm whose input is a 4-tuple $k=(k_1, k_2, k_3, k_4)\in\mathbb{N}^4$ and whose output is an equation involving the $an(i)$. We have written a FORTRAN program that incorporates this algorithm. We leave it to the reader to use (4.3) and Steps 3 and 4 to write such a program. We have found that the set of inputs $k=v(i)(2\leqq i\leqq 37)$, given in Appendix B, yield a system of 36 independent equations in the $an(i)$, as required. In fact a certain sequence of such inputs will yield a certain sequence of equations that can be solved easily using an algebra package like MAPLE. The form of our sequence is shown in Table 1.

We note that each input $v(i_0)$ produces an equation whose left-hand side is $an(i_0)$. Each equation in the output is a linear equation in the $an(i)(1\leqq i\leqq 37)$ and the $an(i)$ that appear on the right-hand side occur as left-hand sides of equations that appear earlier in the output sequence. In other words, the system of equations is *triangular* in shape. Since this paper was first written we have found that this triangularity extends to all root systems. In fact, if the set of inputs has the form

$$\{\gamma\colon \gamma\prec\gamma_0\}\cap L\cap\bar{C},$$

then the corresponding system of equations is triangular with respect to any order that preserves the root order. Here $\prec$ is the usual root order, $\gamma_0\in L\cap\bar{C}$, L is the root lattice and C is the fundamental chamber.

TABLE 1

Input	Output
$k_1 = v(2)$	$an(2) = \dfrac{-a}{(5a+6b+1)}\, an(1)$
$k_2 = v(3)$	$an(3) = \dfrac{-1}{(3a+5b+1)}\{b\, an(1) + 3a\, an(2)\}$
$k_3 = v(4)$	$an(4) = \dfrac{-1}{(3a+4b+1)}\{2(a+b)an(2) + a\, an(3)\}$
$k_4 = v(5)$	$an(5) = \dfrac{-1}{(5a+6b+2)}\{(a+2b)an(2) + 2a\, an(3) + 4(a+b)an(4)\}$
$k_5 = v(7)$	$an(7) = \dfrac{-1}{(5a+6b+2)}\{a\, an(1) + 2a\, an(2) + 6b\, an(3) + 4a\, an(4)\}$
$\vdots$	$\vdots$
$k_{36} = v(37)$	$an(37) = \dfrac{-1}{(6a+8b+6)}\{a\, an(4) + 2a\, an(6) + \cdots + a\, an(36)\}.$

Below we give our complete sequence of inputs:

$$k_1 = v(2),\quad k_2 = v(3),\quad k_3 = v(4),\quad k_4 = v(5),\quad k_5 = v(7),$$
$$k_6 = v(8),\quad k_7 = v(6),\quad k_8 = v(11),\quad k_9 = v(14),\quad k_{10} = v(9),$$
$$k_{11} = v(10),\quad k_{12} = v(12),\quad k_{13} = v(13),\quad k_{14} = v(15),\quad k_{15} = v(16),$$
$$k_{16} = v(17),\quad k_{17} = v(18),\quad k_{18} = v(19),\quad k_{19} = v(20),\quad k_{20} = v(21),$$
$$k_{21} = v(22),\quad k_{22} = v(23),\quad k_{23} = v(24),\quad k_{24} = v(25),\quad k_{25} = v(29),$$
$$k_{26} = v(26),\quad k_{27} = v(27),\quad k_{28} = v(28),\quad k_{29} = v(30),\quad k_{30} = v(31),$$
$$k_{31} = v(32),\quad k_{32} = v(33),\quad k_{33} = v(34),\quad k_{34} = v(35),\quad k_{35} = v(36),$$
$$k_{36} = v(37).$$

We can now complete the proof. From (3.4) and (3.5) we have

(7.1)

$$f'(a+1, b) = 192\, an(1) - 768\, an(2) + \cdots + 192\, an(37)$$

(a complete list of the coefficients is given in Appendix B)

$$= 18432 \frac{(3a+2)(3a+1)(2a+1)(6a+6b+5)(4a+4b+3)}{(3a+4b+3)(3a+5b+3)(5a+6b+5)(2a+3b+1)(3a+5b+2)}$$
$$\cdot \frac{(2a+6b+1)(4a+2b+3)(4a+4b+1)(6a+6b+1)(2a+4b+1)}{(5a+6b+3)(5a+6b+4)(3a+4b+2)(5a+6b+1)(5a+6b+2)}$$
$$\cdot \frac{(4a+2b+1)(2a+2b+1)^2 an(1)}{(3a+4b+1)(3a+5b+1)(2a+3b+2)} \qquad \text{(via MAPLE)}$$
$$= \frac{f(a+1, b)}{f(a, b)} f'(a, b),$$

which is (3.2) as required.

8. Other results. In this section we give other results that have to do with adding roots to the F_4 case of the Macdonald-Morris root system conjecture. Recently [Gal]

we have found a new proof of the G_2 case of the Macdonald-Morris root system conjecture that is solely in terms of integrals. Our proof was motivated by some conjectures of Askey [As] about adding roots to the G_2 case, and is analogous to Aomoto's [Ao] proof of Selberg's integral. We have been able to extend our integral-type proof, mentioned above, to the F_4 case. This proof involves converting equations involving integrals into equations analogous to (4.3) given in Step 3, and then the proof is completed by proceeding as in Steps 4 and 5. The proof given in this paper is more straightforward and direct.

We consider sets of the form $S = T \cup -T$, where T is a subset of the short roots of F_4 (i.e., D_4). We call two such subsets S_1 and S_2 *equivalent* if there is a $w \in W(F_4)$ such that $S_2 = w(S_1)$. This defines an equivalence relation on such sets. By using a FORTRAN program we have found all the equivalence classes, shown in Table 2.

Here the β_k are given in Appendix A. The results for $6 < |S|/2 < 12$ follow easily from the results in the table by taking complements. We denote

$$[x^\alpha] = (1-x^\alpha)(1-x^{-\alpha}). \tag{8.1}$$

We have calculated

$$\text{C.T.} \prod_{\alpha \in T} [x^\alpha] F(x; a, b) \tag{8.2}$$

for all possible subsets T of the short roots of F_4 such that $T \cap -T = \phi$. By Lemma 2.21 it is enough to consider only those T where $S = T \cup -T$ is a representative of an equivalence class. The same FORTRAN program that calculated $f'(a+1, b)$ in terms of the $an(i)$ (see (3.4), (3.5)) was used to calculate (8.2) in terms of the $an(i)$, and hence as a product of a rational function of a and b, and of $f(a, b)$. To our surprise, many of these rational functions factored completely into linear functions. In fact, for each $1 \leqq k \leqq 11$ there exists at least one T with $|T| = k$ such that (8.2) factors completely into linear factors. Most of these linear factors seem to *glue* onto $f(a, b)$ to become factorials. Askey [As] has observed a similar phenomenon in G_2. We do not see how to generalize these results to all root systems. At the very least our results

TABLE 2

$\lvert S\rvert/2$	Number of equivalence classes	Representative of each equivalence class
1	1	$\pm\{\beta_1\}$
2	2	$\pm\{\beta_1, \beta_2\}$, $\pm\{\beta_1, \beta_4\}$
3	4	$\pm\{\beta_1, \beta_2, \beta_3\}$, $\pm\{\beta_1, \beta_2, \beta_4\}$ $\pm\{\beta_1, \beta_4, \beta_5\}$, $\pm\{\beta_2, \beta_4, \beta_5\}$
4	6	$\pm\{\beta_1, \beta_2, \beta_3, \beta_4\}$, $\pm\{\beta_1, \beta_2, \beta_4, \beta_5\}$ $\pm\{\beta_2, \beta_3, \beta_4, \beta_5\}$, $\pm\{\beta_1, \beta_2, \beta_5, \beta_6\}$ $\pm\{\beta_1, \beta_3, \beta_5, \beta_6\}$, $\pm\{\beta_4, \beta_8, \beta_9, \beta_{10}\}$
5	7	$\pm\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$, $\pm\{\beta_1, \beta_2, \beta_3, \beta_5, \beta_6\}$ $\pm\{\beta_1, \beta_2, \beta_4, \beta_5, \beta_6\}$, $\pm\{\beta_1, \beta_3, \beta_4, \beta_5, \beta_6\}$ $\pm\{\beta_3, \beta_4, \beta_5, \beta_6, \beta_8\}$, $\pm\{\beta_1, \beta_2, \beta_3, \beta_7, \beta_8\}$ $\pm\{\beta_1, \beta_4, \beta_8, \beta_9, \beta_{10}\}$
6	9	$\pm\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6\}$, $\pm\{\beta_1, \beta_2, \beta_3, \beta_5, \beta_6, \beta_7\}$ $\pm\{\beta_1, \beta_2, \beta_4, \beta_5, \beta_6, \beta_8\}$, $\pm\{\beta_1, \beta_3, \beta_4, \beta_5, \beta_6, \beta_8\}$ $\pm\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_7, \beta_8\}$, $\pm\{\beta_2, \beta_3, \beta_4, \beta_5, \beta_7, \beta_8\}$ $\pm\{\beta_2, \beta_3, \beta_6, \beta_7, \beta_8, \beta_9\}$, $\pm\{\beta_1, \beta_2, \beta_4, \beta_8, \beta_9, \beta_{10}\}$ $\pm\{\beta_1, \beta_4, \beta_5, \beta_8, \beta_9, \beta_{10}\}$.

seem to indicate that an easier noncomputer proof of F_4 may be possible. The results that only involve linear factors are given below.

For $1 \leqq i_j \leqq 12$ we define

$$[i_1, i_2, \cdots, i_k] = \prod_{j=1}^{k} [\beta_{i_j}]. \tag{8.3}$$

$$\text{C.T.}\,[1]F = 2\frac{(6a+6b+1)}{(5a+6b+1)} f(a, b), \tag{8.4}$$

$$\text{C.T.}\,[1, 2]F = 4\frac{(4a+4b+1)(6a+6b+1)}{(3a+5b+1)(5a+6b+1)} f(a, b), \tag{8.5}$$

$$\text{C.T.}\,[1, 4]F = 2\frac{(4a+4b+1)(6a+6b+1)(5a+10b+2)}{(3a+4b+1)(3a+5b+1)(5a+6b+1)} f(a, b), \tag{8.6}$$

$$\text{C.T.}\,[1, 2, 3]F = 8\frac{(4a+2b+1)(4a+4b+1)(6a+6b+1)}{(3a+4b+1)(3a+5b+1)(5a+6b+1)} f(a, b), \tag{8.7}$$

$$\text{C.T.}\,[1, 4, 5]F = 12\frac{(4a+4b+1)(4a+2b+1)(2a+6b+1)(6a+6b+1)}{(5a+6b+2)(3a+4b+1)(3a+5b+1)(5a+6b+1)} f(a, b), \tag{8.8}$$

$$\text{C.T.}\,[1, 2, 4, 5]F = 8\frac{(2a+6b+1)(4a+2b+1)(4a+4b+1)(7a+8b+3)}{(2a+3b+1)(5a+6b+2)(3a+4b+1)(3a+5b+1)} \cdot \frac{(6a+6b+1)}{(5a+6b+1)} f(a, b), \tag{8.9}$$

$$\text{C.T.}\,[4, 8, 9, 10]F = 32\frac{(3a+1)(6a+6b+1)(4a+4b+1)(4a+2b+1)}{(5a+6b+2)(3a+4b+1)(3a+5b+1)(5a+6b+1)} f(a, b), \tag{8.10}$$

$$\text{C.T.}\,[1, 4, 8, 9, 10]F = 48\frac{(3a+1)(2a+2b+1)(4a+2b+1)(6a+6b+1)}{(5a+6b+3)(2a+3b+1)(5a+6b+2)(3a+4b+1)} \cdot \frac{(4a+4b+1)(7a+12b+4)}{(3a+5b+1)(5a+6b+1)} f(a, b), \tag{8.11}$$

$$\text{C.T.}\,[1, 2, 4, 5, 6, 8]F = 288\frac{(3a+1)(6a+6b+1)(2a+6b+1)(2a+2b+1)}{(5a+6b+3)(3a+5b+2)(2a+3b+1)(5a+6b+2)} \cdot \frac{(4a+4b+1)(2a+4b+1)(4a+2b+1)}{(3a+4b+1)(3a+5b+1)(5a+6b+1)} f(a, b), \tag{8.12}$$

$$\text{C.T.}\,[1, 2, 7, 9, 10, 11, 12]F = 384\frac{(3a+1)(2a+2b+1)(2a+4b+1)(2a+6b+1)(5a+5b+3)}{(3a+4b+2)(5a+6b+3)(3a+5b+2)(2a+3b+1)(5a+6b+2)} \cdot \frac{(4a+4b+1)(6a+6b+1)(4a+2b+1)}{(3a+4b+1)(3a+5b+1)(5a+6b+1)} f(a, b), \tag{8.13}$$

(8.14) C.T. $[1, 2, 3, 5, 6, 7, 11, 12]F$

$$=1152\frac{(3a+2)(3a+1)(4a+4b+3)(2a+2b+1)(4a+6b+3)}{(5a+6b+4)(3a+4b+2)(5a+6b+3)(3a+5b+2)(2a+3b+1)}$$
$$\cdot\frac{(4a+4b+1)(4a+2b+1)(6a+6b+1)(4a+2b+3)}{(5a+6b+2)(3a+4b+1)(3a+5b+1)(5a+6b+1)}f(a, b),$$

(8.15) C.T. $[1, 6, 7, 8, 9, 10, 11, 12]F$

$$=384\frac{(3a+1)(4a+2b+1)(2a+4b+1)(4a+4b+3)(2a+2b+1)}{(5a+6b+4)(3a+4b+2)(5a+6b+3)(3a+5b+2)(2a+3b+1)}$$
$$\cdot\frac{(6a+6b+1)(2a+6b+1)(4a+4b+1)(13a+10b+7)}{(5a+6b+2)(3a+4b+1)(3a+5b+1)(5a+6b+1)}f(a, b),$$

(8.16) C.T. $[2, 4, 7, 8, 9, 10, 11, 12]F$

$$=768\frac{(3a+1)(6a+6b+1)(2a+6b+1)(4a+2b+1)(4a+4b+3)}{(5a+6b+4)(3a+4b+2)(5a+6b+3)(3a+5b+2)(2a+3b+1)}$$
$$\cdot\frac{(4a+4b+1)(2a+4b+1)(2a+2b+1)(7a+4b+4)}{(5a+6b+2)(3a+4b+1)(3a+5b+1)(5a+6b+1)}f(a, b),$$

(8.17) C.T. $[1, 3, 6, 7, 8, 9, 10, 11, 12]F$

$$=2304\frac{(3a+1)(2a+4b+1)(6a+6b+5)(4a+4b+3)(2a+6b+1)}{(5a+6b+4)(3a+5b+3)(3a+4b+2)(5a+6b+3)(3a+5b+2)}$$
$$\cdot\frac{(4a+4b+1)(6a+6b+1)(4a+2b+1)(2a+2b+1)^2}{(2a+3b+1)(5a+6b+2)(3a+4b+1)(3a+5b+1)(5a+6b+1)}$$
$$\cdot f(a, b),$$

(8.18) C.T. $[3, 5, 6, 7, 8, 9, 10, 11, 12]F$

$$=768\frac{(3a+1)(4a+2b+1)(2a+6b+1)(6a+6b+1)(6a+6b+5)}{(5a+6b+4)(3a+5b+3)(3a+4b+2)(5a+6b+3)(3a+5b+2)}$$
$$\cdot\frac{(4a+4b+1)(4a+4b+3)(2a+4b+1)(2a+2b+1)(7a+4b+4)}{(2a+3b+1)(5a+6b+2)(3a+4b+1)(3a+5b+1)(5a+6b+1)}$$
$$\cdot f(a, b),$$

(8.19) C.T. $[4, 5, 6, 7, 8, 9, 10, 11, 12]F$

$$=2304\frac{(3a+2)(3a+1)(6a+5+6b)(4a+4b+3)(6a+6b+1)}{(5a+6b+5)(5a+6b+4)(3a+4b+2)(5a+6b+3)(3a+5b+2)}$$
$$\cdot\frac{(4a+4b+1)(2a+6b+1)(2a+2b+1)(4a+2b+3)(4a+2b+1)}{(2a+3b+1)(5a+6b+2)(3a+4b+1)(3a+5b+1)(5a+6b+1)}$$
$$\cdot f(a, b),$$

(8.20) C.T. $[2, 3, 5, 6, 7, 8, 9, 10, 11, 12]F$

$$=2304\frac{(3a+1)(2a+4b+1)(6a+6b+5)(4a+2b+1)(2a+6b+1)}{(2a+3b+2)(5a+6b+4)(3a+5b+3)(3a+4b+2)(5a+6b+3)}$$
$$\cdot\frac{(6a+6b+1)(4a+4b+1)(4a+2b+3)(4a+4b+3)(2a+2b+1)^2}{(3a+5b+2)(2a+3b+1)(5a+6b+2)(3a+4b+1)(3a+5b+1)}$$
$$\cdot\frac{f(a, b)}{(5a+6b+1)},$$

(8.21) C.T. $[3, 4, 5, 6, 7, 8, 9, 10, 11, 12]F$

$$=4608\frac{(3a+2)(3a+1)(6a+6b+1)(4a+4b+1)(2a+6b+1)}{(5a+6b+5)(5a+6b+4)(3a+5b+3)(3a+4b+2)(5a+6b+3)}$$
$$\cdot\frac{(4a+2b+1)(2a+4b+1)(6a+6b+5)(4a+2b+3)(2a+2b+1)}{(3a+5b+2)(2a+3b+1)(5a+6b+2)(3a+4b+1)(3a+5b+1)}$$
$$\cdot\frac{(4a+4b+3)}{(5a+6b+1)}f(a, b),$$

(8.22) C.T. $[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]F$

$$=3072\frac{(3a+2)(3a+1)(6a+6b+5)(2a+4b+1)(6a+6b+1)}{(3a+5b+1)(5a+6b+1)(3a+4b+2)(5a+6b+3)(3a+5b+2)}$$
$$\cdot\frac{(4a+4b+1)(4a+2b+1)(2a+6b+1)(4a+4b+3)(4a+2b+3)}{(2a+3b+1)(5a+6b+2)(3a+4b+1)(5a+6b+5)(5a+6b+4)}$$
$$\cdot\frac{(2a+2b+1)^2(7a+9b+7)}{(3a+5b+3)(2a+3b+2)(3a+4b+3)}f(a, b).$$

It is a little unsettling that not all of the results above can be written as factorials. Since this paper was first written we have found nicer results. In fact, if we restrict ourselves to subsets of positive roots, there is a chain of subsets in which each corresponding constant term formula can be written as a product of factorials. This chain seems to be related to the root order, but we have been unable to generalize to other root systems.

Appendix A. We write the roots of $\Phi^{(2)}(F_4)$ as $\mathbb{Z}$-linear combinations of $\alpha_i (1\leqq i\leqq 4)$ defined in (2.6):

$$\beta_1=\alpha_1=e_1-e_2,\quad \beta_5=\alpha_1+\alpha_4=e_1-e_3,\quad \beta_8=\alpha_1+\alpha_2+\alpha_4=e_1+e_4,$$
$$\beta_2=\alpha_2=e_3+e_4,\quad \beta_6=\alpha_2+\alpha_4=e_2+e_4,\quad \beta_9=\alpha_1+\alpha_3+\alpha_4=e_1-e_4,$$
$$\beta_3=\alpha_3=e_3-e_4,\quad \beta_7=\alpha_3+\alpha_4=e_2-e_4,\quad \beta_{10}=\alpha_2+\alpha_3+\alpha_4=e_2+e_3,$$
$$\beta_4=\alpha_4=e_2-e_3,$$
$$\beta_{11}=\alpha_1+\alpha_2+\alpha_3+\alpha_4=e_1+e_3,$$
$$\beta_{12}=\alpha_1+\alpha_2+\alpha_3+2\alpha_4=e_1+e_2,$$
$$\alpha_1+\alpha_2=e_1-e_2+e_3+e_4,\qquad \alpha_1-\alpha_2=e_1-e_2-e_3-e_4,$$
$$\alpha_1+\alpha_3=e_1-e_2+e_3-e_4,\qquad \alpha_1-\alpha_3=e_1-e_2-e_3+e_4,$$
$$\alpha_2+\alpha_3=2e_3,\qquad \alpha_2-\alpha_3=2e_4,$$
$$\alpha_1+\alpha_2+2\alpha_4=e_1+e_2-e_3+e_4,\qquad 2\alpha_1+\alpha_2+\alpha_3+2\alpha_4=2e_1,$$
$$\alpha_1+\alpha_3+2\alpha_4=e_1+e_2-e_3-e_4,\qquad \alpha_1+2\alpha_2+\alpha_3+2\alpha_4=e_1+e_2+e_3+e_4,$$
$$\alpha_2+\alpha_3+2\alpha_4=2e_2,\qquad \alpha_1+\alpha_2+2\alpha_3+2\alpha_4=e_1+e_2+e_3-e_4.$$

Appendix B. The vectors $v(i)(1 \leqq i \leqq 37)$ that appear in (3.5) are listed below:

$$v(1)=(0,0,0,0), \quad v(14)=(2,2,2,2), \quad v(27)=(5,3,3,1),$$

$$v(2)=(1,1,0,0), \quad v(15)=(4,2,2,0), \quad v(28)=(5,4,1,0),$$

$$v(3)=(1,1,1,1), \quad v(16)=(4,4,0,0), \quad v(29)=(4,4,2,2),$$

$$v(4)=(2,1,1,0), \quad v(17)=(3,3,2,2), \quad v(30)=(5,4,2,1),$$

$$v(5)=(2,2,1,1), \quad v(18)=(4,2,2,2), \quad v(31)=(6,2,2,2),$$

$$v(6)=(3,2,1,0), \quad v(19)=(4,3,2,1), \quad v(32)=(6,3,2,1),$$

$$v(7)=(2,2,0,0), \quad v(20)=(4,4,1,1), \quad v(33)=(6,3,3,0),$$

$$v(8)=(3,1,1,1), \quad v(21)=(5,2,2,1), \quad v(34)=(5,5,0,0),$$

$$v(9)=(3,2,2,1), \quad v(22)=(5,3,1,1), \quad v(35)=(5,5,1,1),$$

$$v(10)=(3,3,1,1), \quad v(23)=(5,3,2,0), \quad v(36)=(6,4,1,1),$$

$$v(11)=(3,3,0,0), \quad v(24)=(3,3,3,3), \quad v(37)=(6,4,2,0).$$

$$v(12)=(4,2,1,1), \quad v(25)=(4,3,3,2),$$

$$v(13)=(4,3,1,0), \quad v(26)=(5,3,2,2),$$

The complete version of (3.5) is given below:

$$\begin{aligned}\sum_{\alpha \in S} a'_\alpha x^\alpha = {} & 192x^{v(1)}-768x^{v(2)}+576x^{v(3)}+960x^{v(4)}-1152x^{v(5)}+2688x^{v(6)} \\ & -576x^{v(7)}-1152x^{v(8)}-576x^{v(10)}-576x^{v(11)}-1152x^{v(12)} \\ & +1152x^{v(13)}+576x^{v(14)}-192x^{v(15)}-192x^{v(16)}-1152x^{v(17)} \\ & +2304x^{v(18)}-1152x^{v(21)}-1152x^{v(22)}+1152x^{v(23)}+576x^{v(24)} \\ & -1728x^{v(25)}+2304x^{v(26)}-1152x^{v(27)}+384x^{v(28)}+576x^{v(29)} \\ & -1152x^{v(30)}-576x^{v(31)}+1152x^{v(32)}-192x^{v(33)}-192x^{v(34)} \\ & +576x^{v(35)}-576x^{v(36)}+192x^{v(37)}.\end{aligned}$$

Acknowledgments. I thank the following: Richard Askey for encouraging me to tackle F_4, Kevin Kadell for pointing out the approach of working with Laurent polynomials rather than with integrals, and Dennis Stanton for suggesting many improvements and for helping me with Fig. 1. It should be pointed out that Kevin Kadell has obtained independently some of the results of this paper. In particular he has calculated the values of $an(2)$ and $an(3)$ given in Table 1 in § 7.

Note added in proof. Robert Gustafson (*A generalization of Selberg's beta integral*, preprint) has proved q-Macdonald-Morris for the affine root systems of types $S(C_l)$ and $S(C_l)^\vee$.

REFERENCES

[An] G. E. ANDREWS, *Problems and prospects for basic hypergeometric functions*, in Theory and Applications of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 191-224.

[Ao] K. AOMOTO, *Jacobi polynomials associated with Selberg's integral*, SIAM J. Math. Anal., 18 (1987), pp. 545-549.

[As] R. ASKEY, *Integration and computers*, in Proc. Computer Algebra Conference, D. Chudnowsky, G. Chudnowsky, and R. Jenks, eds., to appear.

[Ba] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, London, New York, 1935; reprinted, Hafner, New York, 1964.

[Bo] N. BOURBAKI, *Groupes et algèbres de Lie*, Chaps. 4-6, Hermann, Paris, 1968.

[Ca] R. W. CARTER, *Simple Groups of Lie Type*, John Wiley, London, New York, 1972.

[Co] A. G. CONSTANTINE, *Some noncentral distribution problems in multivariate analysis*, Ann. Math. Statist., 34 (1963), pp. 1270-1285.

[D] F. J. DYSON, *Statistical theory of the energy levels of complex systems*: I, J. Math. Phys., 3 (1962), pp. 140-156.

[E] R. EVANS, *Character sum analogues of constant term identities for root systems*, Israel J. Math., 46 (1983), pp. 189-196.

[Ga1] F. G. GARVAN, *A beta integral associated with the root system G_2*, SIAM J. Math. Anal., 19 (1988), pp. 1462-1474.

[Ga2] ———, *Some Macdonald-Mehta integrals by brute force*, in q-Series and Partitions, D. Stanton, ed., IMA Vol. in Math. Appl. 18, Springer-Verlag, Berlin, New York, 1989, pp. 77-98.

[Ga3] ———, *Progress on the Macdonald conjectures*, preprint.

[Go] I. J. GOOD, *Short proof of a conjecture of Dyson*, J. Math. Phys., 11 (1970), p. 1884.

[Gu] J. GUNSON, *Proof of a conjecture of Dyson in the statistical theory of energy levels*, J. Math. Phys., 3 (1962), pp. 752-753.

[Hab1] L. HABSIEGER, *Une q-intégrale de Selberg et Askey*, SIAM J. Math. Anal., 19 (1988), pp. 1475-1489; summarized in C.R. Acad. Sci., 302 (1986), pp. 615-617.

[Hab2] ———, *La q-conjecture de Macdonald-Morris pour G_2*, C.R. Acad. Sci., Paris Sér. I Math., 303 (1986), pp. 211-213.

[Han1] P. HANLON, *On the decomposition of the tensor algebra of the classical Lie algebras*, Adv. in Math., 56 (1985), pp. 238-282.

[Han2] ———, *Cyclic homology and the Macdonald conjectures*, Invent. Math., 86 (1986), pp. 131-159.

[He] G. J. HECKMAN, *Root systems and hypergeometric functions* II, Compositio Math., 64 (1987), pp. 353-373.

[Hu] J. H. HUMPHREYS, *Introduction to Lie Algebras and Representation Theory*, Springer-Verlag, Berlin, New York, 1972.

[Jac] H. JACK, *A class of symmetric polynomials with a parameter*, Proc. Roy. Soc. Edinburgh Sect. A, 69 (1969-70), pp. 1-17.

[Jam1] A. T. JAMES, *Zonal polynomials of the real positive definite matrices*, Ann. Math., 74 (1961), pp. 456-469.

[Jam2] ———, *Distribution of matrix variables and latent roots derived from normal samples*, Ann. Math. Statist., 35 (1964), pp. 475-501.

[Jam3] ———, *Calculation of zonal polynomial coefficients by use of the Laplace-Beltrami operator*, Ann. Math. Statist., 39 (1968), pp. 1711-1718.

[K1] K. W. J. KADELL, *A proof of Askey's conjectured q-analog of Selberg's integral and a conjecture of Morris*, SIAM J. Math. Anal., 19 (1988), pp. 969-986.

[K2] ———, *A proof of the q-Macdonald-Morris conjecture for BC_n*, preprint.

[K3] ———, *The q-Selberg-Jack polynomials*, preprint.

[Ma1] I. G. MACDONALD, *Affine root systems and Dedekind's η-function*, *Invent. Math.*, 15 (1972), pp. 91-143.

[Ma2] ———, *Some conjectures for root systems and finite reflection groups*, SIAM J. Math. Anal., 13 (1982), pp. 988-1007.

[Ma3] ———, *Symmetric Functions and Hall Polynomials*, Second edition, to appear.

[Me] M. L. MEHTA, *Random Matrices*, Academic Press, New York, London, 1967.

[M-D] M. L. MEHTA AND F. J. DYSON, *Statistical theory of the energy levels of complex systems*: V, J. Math. Phys., 4 (1963), pp. 713-719.

[Mo] W. G. MORRIS, *Constant term identities for finite and affine root systems*, Ph.D. thesis, Univ. of Wisconsin, Madison, WI, 1982.

[O1] E. M. OPDAM, *Generalized hypergeometric functions associated with root systems*, Ph.D. thesis, Leiden, the Netherlands, 1988.

[O2] E. M. OPDAM, *Some applications of hypergeometric shift operators*, preprint.

[Re1] A. REGEV, *Asymptotic values for degrees associated with strips of Young diagrams*, Adv. in Math., 41 (1981), pp. 115-136.

[Re2] ———, *Combinatorial sums, identities and trace identities of the* 2×2 *matrices*, Adv. in Math., 46 (1982), pp. 230-240.

[Ri] D. RICHARDS, *Some extensions of Selberg's and related integrals, with numerous applications*, Presented at q-Series Workshop, Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, MN, March 1988.

[Se] A. SELBERG, *Bermerkniger om et multiplet integral*, Norsk. Mat. Tidsskr., 26 (1944), pp. 71-78.

[St] R. P. STANLEY, *Some combinatorial properties of Jack symmetric functions*, preprint.

[T] E. C. TITCHMARSH, *The Theory of Functions*, Oxford University Press, Oxford, London, 1939.

[W] K. WILSON, *Proof of a conjecture of Dyson*, J. Math. Phys., 3 (1962), pp. 1040-1043.

[Z1] D. ZEILBERGER, *A combinatorial proof of Dyson's conjecture*, Discrete Math., 41 (1982), pp. 317-321.

[Z2] ———, *A proof of the* G_2 *case of Macdonald's root system-Dyson conjecture*, SIAM J. Math. Anal., 18 (1987), pp. 880-883.

[Z3] ———, *A unified approach to Macdonald's root-system conjectures*, SIAM J. Math. Anal., 19 (1988), pp. 987-1013.

[Z-B] D. ZEILBERGER AND D. BRESSOUD, *A proof of the Andrews' q-Dyson conjecture*, Discrete Math., 54 (1985), pp. 201-224.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 823–836, July 1990
©1990 Society for Industrial and Applied Mathematics

DERIVATION OF THE DOUBLE POROSITY MODEL OF SINGLE PHASE FLOW VIA HOMOGENIZATION THEORY*

TODD ARBOGAST†, JIM DOUGLAS, JR.†, AND ULRICH HORNUNG‡

Abstract. A general form of the double porosity model of single phase flow in a naturally fractured reservoir is derived from homogenization theory. The microscopic model consists of the usual equations describing Darcy flow in a reservoir, except that the porosity and permeability coefficients are highly discontinuous. Over the matrix domain, the coefficients are scaled by a parameter ϵ representing the size of the matrix blocks. This scaling preserves the physics of the flow in the matrix as ϵ tends to zero. An effective macroscopic limit model is obtained that includes the usual Darcy equations in the matrix blocks and a similar equation for the fracture system that contains a term representing a source of fluid from the matrix. The convergence is shown by extracting weak limits in appropriate Hilbert spaces. A dilation operator is utilized to see the otherwise vanishing physics in the matrix blocks as ϵ tends to zero.

Key words. porous medium, double porosity, fractured reservoir, homogenization

AMS(MOS) subject classification. 76S05

1. Introduction.

It has long been known that the porous rock that composes a petroleum reservoir may contain many cracks or *fractures*. A naturally fractured reservoir is one that has throughout its extent many interconnected fracture planes. For over 30 years it has been known that flow in such reservoirs is not like that in unfractured reservoirs [16]. Instead, the flow acts as if the reservoir possessed two porous structures, one associated to the system of fractures, and the other associated to the porous rock (the *matrix*). This double porosity/permeability concept has been used to model the flow of a single component in a single phase within a naturally fractured reservoir since around 1960 [5], [14], [18], [21].

More recently, a general form of the double porosity/permeability model has been described [3], [10]. The earlier models can be considered as approximations to this more general model [3]. It was derived on physical grounds under the main assumption that the fluid pressure (or, equivalently, density) is uniform at the surface of each matrix block. Herein we will derive this general model from the point of view of homogenization theory [6], [17]. It will be seen that the model is in some sense the limit of a family of microscopic models in which the sizes of the matrix blocks tend to zero (hence, the fluid indeed becomes uniform at the surfaces of the blocks).

A straightforward homogenization of the entire reservoir would yield a single porosity model with some average permeability [1], [7]. This would be quite inadequate since two very distinct porous structures are present in the reservoir and their interaction has a strong influence on the flow characteristics. This interaction is a fine structure process whose effect only must be homogenized; the process itself must be retained on a microscopic level.

*Received by the editors January 16, 1989; accepted for publication August 1, 1989. This work was supported in part by the National Science Foundation.

†Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

‡Scientific Hornung Institute (SCHI), P.O. Box 1222, D-8014 Neubiberg, Federal Republic of Germany.

We will use a variant of the homogenization technique that was used by Hornung and Jäger to describe catalytic reactions in a porous medium [12]. There, the driving mechanism was the chemical process of catalytic reactions that takes place on a microscopic scale. The overall (or homogenized) behavior of the system could be obtained only when a careful modeling of these microscopic reactions was maintained as the medium was homogenized. See also [11] for a similar situation involving displacement in mobile and immobile water and [20] for an example from chromatography. This technique is also discussed in a formal sense in a related paper where Arbogast, Douglas, and Hornung consider two-component flows in naturally fractured reservoirs [4]; see also [9]. An independent study of diffusion problems in fractured porous media obtained by homogenization is given by Hornung and Showalter [13].

The remainder of the paper is as follows. In the next section, we will write down the equations that describe the microscopic nature of the flow in a naturally fractured reservoir. We will scale the equations by a parameter ϵ, where ϵ is the size of the matrix blocks. This places the model in a series of problems from which, as ϵ tends to zero, we obtain our homogenized macroscopic model, presented in §3. In the final section, we prove that the solutions of the microscopic model converge weakly to those of the macroscopic model in appropriate Hilbert spaces and describe some of the mathematical properties of the limit model.

2. The microscopic model. We consider the reservoir $\Omega \subset \mathbb{R}^3$ to be a bounded, two-connected domain with a periodic structure. More precisely, Ω is a union of disjoint parallelepiped cell domains congruent to a standard one $\mathcal{Q}$:

$$\overline{\Omega} = \bigcup_{c \in \mathcal{A}} (\overline{\mathcal{Q}} + c), \qquad (\mathcal{Q} + c_1) \cap (\mathcal{Q} + c_2) = \emptyset \quad \text{whenever } c_1 \neq c_2 \in \mathcal{A},$$

where $\mathcal{A}$ is an appropriate finite lattice of translations containing the origin and the overline denotes closure (see Fig. 1). The cell $\mathcal{Q}$ can be decomposed into three pieces, a compactly contained, two-connected domain $\mathcal{Q}_m$ representing the matrix block part, the surrounding connected fracture domain $\mathcal{Q}_f$, and a smooth internal boundary piece $\partial\mathcal{Q}_m$ (see Fig. 2).

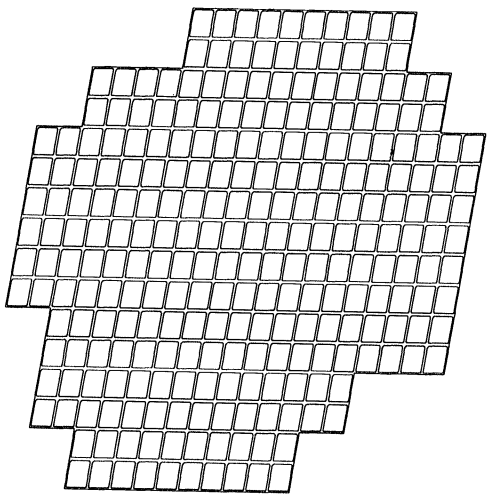


FIG. 1. *The reservoir Ω, depicting its periodic structure.*

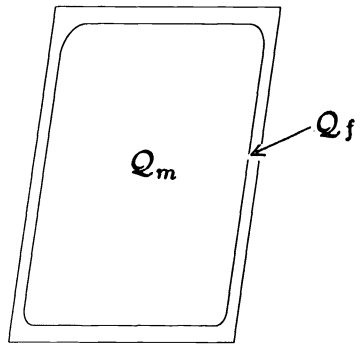


FIG. 2. *The standard cell Q.*

To homogenize the reservoir, we will let tend to zero the (linear) size ϵ of the cells ($\epsilon = 1$ in the microscopic model). Extend $\mathcal{A}$ to an infinite lattice $\mathcal{A}'$. For $\epsilon > 0$, let the fracture and matrix domains be denoted, respectively, by

$$\Omega_f^\epsilon = \Omega \cap \bigcup_{c \in \mathcal{A}'} \epsilon(\mathcal{Q}_f + c) \qquad \text{and} \qquad \Omega_m^\epsilon = \Omega \cap \bigcup_{c \in \mathcal{A}'} \epsilon(\mathcal{Q}_m + c).$$

To avoid unimportant technicalities relating to the boundary of Ω, assume that the ϵ's form a sequence for which $\partial\Omega \subset \partial\Omega_f^\epsilon$.

Let us define some notation and make some physical assumptions before setting up the microscopic model. Denote by $\rho^\epsilon(x,t)$ and $\sigma^\epsilon(x,t)$ the density of the fluid in Ω_f^ϵ and Ω_m^ϵ, respectively. Assume that the fluid is a liquid of viscosity μ and constant compressibility c; that is, the pressure p and the density satisfy the equations of state:

$$d\rho = c\rho\, dp \qquad \text{and} \qquad d\sigma = c\sigma\, dp.$$

Of course we will assume that the fluid flows according to Darcy's law in the matrix Ω_m^ϵ, where we will let $k^\epsilon(x) = k(x/\epsilon)$ and $\phi^\epsilon(x) = \phi(x/\epsilon)$ denote the porosity and (possibly tensor) permeability, respectively. These quantities should be periodic of period $\mathcal{Q}$ (reflecting the periodicity of the matrix blocks over Ω—more generally we could assume that the fixed properties of the matrix are of the form $\psi^\epsilon(x) = \psi(x, x/\epsilon)$, varying over the reservoir in the first argument and periodic in the second).

We will also assume Darcy's law is valid in the fractures Ω_f^ϵ. Clearly this is not strictly correct; however, this has been done in the petroleum engineering literature [14], [18], evidently by considering the fractures to be partially filled with rock debris. In any case, Darcy's law should hold as ϵ tends to zero [3], [5], [16], [19], [21]; our main interest lies in determining the correct form of the interaction between the matrix and fracture systems. So, let $\Phi^*(x)$ (≈ 1) and $K^*(x)$ (very large) denote the porosity and scalar permeability of the fracture domain, extended over all of Ω. These four quantities are uniformly positive (k being symmetric, uniformly positive-definite), and each is assumed smooth and bounded.

Finally, choose some smooth and bounded reference density functions $\rho_{\text{ref}}(x)$ and $\sigma_{\text{ref}}^\epsilon(x) = \sigma_{\text{ref}}(x/\epsilon)$, where $\sigma_{\text{ref}}(x)$ is periodic as above. To linearize the equations, we

will approximate the effects of gravity as follows:

$$(\rho^\epsilon)^2 \approx \rho_{\text{ref}}(2\rho^\epsilon - \rho_{\text{ref}}), \qquad (\sigma^\epsilon)^2 \approx \sigma^\epsilon_{\text{ref}}(2\sigma^\epsilon - \sigma^\epsilon_{\text{ref}}).$$

The flow in the two domains is then described by conservation of mass combined with Darcy's law. Using our assumptions, in the fracture domain we have

$$\Phi^*\rho^\epsilon_t - \nabla \cdot \left\{ \frac{K^*}{\mu c}\left[\nabla\rho^\epsilon + cg\rho_{\text{ref}}(2\rho^\epsilon - \rho_{\text{ref}})\right]\right\} = f \quad \text{for } x \in \Omega^\epsilon_f, \quad t > 0, \tag{2.1}$$

$$\left\{ \frac{K^*}{\mu c}\left[\nabla\rho^\epsilon + cg\rho_{\text{ref}}(2\rho^\epsilon - \rho_{\text{ref}})\right]\right\} \cdot \nu = \epsilon\left\{ \frac{k^\epsilon}{\mu c}\left[\epsilon\nabla\sigma^\epsilon + cg\sigma^\epsilon_{\text{ref}}(2\sigma^\epsilon - \sigma^\epsilon_{\text{ref}})\right]\right\} \cdot \nu \quad \text{for } x \in \partial\Omega^\epsilon_m, \quad t > 0, \tag{2.2}$$

$$\rho^\epsilon = \rho_{\text{init}} \quad \text{for } x \in \Omega^\epsilon_f, \quad t = 0; \tag{2.3}$$

here, the subscript t denotes partial differentiation in time, g is the gravitational constant vector, $f(x,t)$ represents external sources/sinks, $\nu(x)$ is the outer unit normal (to $\partial\Omega^\epsilon_m$, in this case), and $\rho_{\text{init}}(x)$ is the specified initial density.

Similarly in the matrix domain,

$$\phi^\epsilon\sigma^\epsilon_t - \epsilon\nabla \cdot \left\{ \frac{k^\epsilon}{\mu c}\left[\epsilon\nabla\sigma^\epsilon + cg\sigma^\epsilon_{\text{ref}}(2\sigma^\epsilon - \sigma^\epsilon_{\text{ref}})\right]\right\} = f \quad \text{for } x \in \Omega^\epsilon_m, \quad t > 0, \tag{2.4}$$

$$\sigma^\epsilon = \rho^\epsilon \quad \text{for } x \in \partial\Omega^\epsilon_m, \quad t > 0, \tag{2.5}$$

$$\sigma^\epsilon = \rho_{\text{init}} \quad \text{for } x \in \Omega^\epsilon_m, \quad t = 0. \tag{2.6}$$

The two boundary conditions (2.2) and (2.5) represent conservation of mass flux and continuity of pressure, respectively, between the two domains. We should also assume a no-flow Neumann condition on the fracture density on $\partial\Omega \subset \partial\Omega^\epsilon_f$:

$$\left\{ \frac{K^*}{\mu c}\left[\nabla\rho^\epsilon + cg\rho_{\text{ref}}(2\rho^\epsilon - \rho_{\text{ref}})\right]\right\} \cdot \nu = 0 \quad \text{for } x \in \partial\Omega, \quad t > 0. \tag{2.7}$$

We should make some remarks on the ϵ scaling factors. They can be viewed as coming from a dimensional analysis of the equations for an individual matrix block; they provide the correct scaling for the flow as the block size shrinks. That is, the form of the matrix equations is preserved on the standard cell independently of the value of ϵ, thereby giving a double porosity model in the limit. This will be seen explicitly in §4. Alternatively, we may say that the flow between the matrix and fractures is conserved in some sense, as it is prevented from degenerating or blowing up as $\epsilon \to 0$ [4]. Essentially, the matrix permeability has been scaled by ϵ^2, whereas the gravitational term has been compensated by ϵ^{-1}. We might instead decide not to compensate gravity, in which case the macroscopic model would have no gravitational terms in the matrix equations.

3. The macroscopic model. We will show in the next section that, as the scaling parameter ϵ tends to zero, the microscopic model converges in some sense to the model described below.

We must first define some symbols. Define the macroscopic fracture porosity as

$$\Phi(x) = \frac{|\mathcal{Q}_f|}{|\mathcal{Q}|}\Phi^*(x), \tag{3.1}$$

where $|\cdot|$ denotes the volume of the set. Let us define the auxiliary functions $\omega_j(y)$, $j = 1, 2, 3$, periodic of period $\mathcal{Q}$, as the solutions modulo constants of

$$\Delta\omega_j = 0 \quad \text{for } y \in \mathcal{Q}_f, \tag{3.2}$$

$$\nabla\omega_j \cdot \nu = -e_j \cdot \nu = -\nu_j \quad \text{for } y \in \partial\mathcal{Q}_m, \tag{3.3}$$

where e_j is the unit vector in the jth direction. Now we can define the effective macroscopic fracture system permeability tensor K componentwise as

$$K_{ij}(x) = \frac{1}{|\mathcal{Q}|} K^*(x) \left\{ |\mathcal{Q}_f| \delta_{ij} + \int_{\mathcal{Q}_f} \partial_i \omega_j \, dy \right\}, \tag{3.4}$$

where δ_{ij} is the Kronecker symbol and $\partial_i = \partial/\partial y_i$.

With f_m being defined in (3.11) below, the macroscopic fracture density $\rho(x,t)$ satisfies

$$\Phi\rho_t - \nabla \cdot \left\{ \frac{K}{\mu c} \left[\nabla\rho + cg\rho_{\text{ref}}(2\rho - \rho_{\text{ref}}) \right] \right\} = f + f_m \quad \text{for } x \in \Omega, \quad t > 0, \tag{3.5}$$

$$\left\{ \frac{K}{\mu c} \left[\nabla\rho + cg\rho_{\text{ref}}(2\rho - \rho_{\text{ref}}) \right] \right\} \cdot \nu = 0 \quad \text{for } x \in \partial\Omega, \quad t > 0, \tag{3.6}$$

$$\rho = \rho_{\text{init}} \quad \text{for } x \in \Omega, \quad t = 0. \tag{3.7}$$

As ϵ tends to zero, we obtain an infinite number of matrix blocks, one for each x. Hence, for each $x \in \Omega$, we have a matrix density function $\sigma(x,y,t)$, which is determined from

$$\phi(y)\sigma_t - \nabla_y \cdot \left\{ \frac{k(y)}{\mu c} \left[\nabla_y\sigma + cg\sigma_{\text{ref}}(y)(2\sigma - \sigma_{\text{ref}}(y)) \right] \right\} = f(x,t) \tag{3.8}$$

$$\text{for } y \in \mathcal{Q}_m, \quad t > 0,$$

$$\sigma = \rho(x,t) \quad \text{for } y \in \partial\mathcal{Q}_m, \quad t > 0, \tag{3.9}$$

$$\sigma = \rho_{\text{init}}(x) \quad \text{for } y \in \mathcal{Q}_m, \quad t = 0, \tag{3.10}$$

where ∇_y is the gradient with respect to the y variable. Finally, the matrix source term f_m is defined by

$$f_m(x,t) = -\frac{1}{|\mathcal{Q}|} \int_{\mathcal{Q}_m} \phi(y)\sigma_t(x,y,t) \, dy. \tag{3.11}$$

Except for some minor differences, this is exactly the general model originally described in [3] and [10]. Two obvious but unimportant differences are that in the original formulation of the model, the matrix equations do not contain either gravity or external sources; these terms could easily be included. In any case, they are of little mathematical consequence. The interesting differences are associated to the homogenization process itself. The original formulation (and, in fact, the actual problem) had a finite number of finite-sized blocks; whereas in the homogenized version of the model, there is a continuum of blocks having no size, since ϵ tended to zero. This gives rise to subtle differences in the form of the matrix boundary and initial conditions (3.9) and (3.10) as well as a difference in the definition of the matrix source term (3.11). However, there is no difference of any real substance. The boundary and initial conditions of the matrix problems of both formulations are constant with respect to the space variables of the block. Of course in its original form, some representative value of the fracture density over the finite-sized block needed to be used. The matrix source terms of both formulations represent the total average flow out of the matrix

blocks; this flow needed to be placed explicitly over the entire extent of the finite-sized block in the original version. In practice, the two formulations of the model would be discretized in essentially the same way, since we would need to restrict to a finite number of matrix block problems.

4. The convergence of the homogenization process. We begin by defining some notation. Denote the Sobolev space of functions with derivatives of order m in L^p by $W^{m,p}$, and set $H^m = W^{m,2}$. Let H_0^1 denote the closure in H^1 of C_0^∞, the infinitely differentiable functions with compact support. Denote by $W^{m,p}(W^{n,q})$ the space of $W^{n,q}$-valued functions in $W^{m,p}$. Let $J = (0,T]$, $T > 0$, be the time interval of interest. We will make use of several bilinear integration forms. Denote integration in space over the domain $*$ by $(\cdot,\cdot)_*$, where $*$ is some subset of Ω, $\mathcal{Q}$, or $\Omega \times \mathcal{Q}$, and where $*$ determines which of the variables to integrate over. Also, denote integration in space over $*$ and in time over J by $\langle \cdot,\cdot \rangle_*$. Let us simplify the notation a bit by setting

$$\Lambda(x) = \frac{K^*(x)}{\mu c}, \quad \lambda^\epsilon(x) = \frac{k^\epsilon(x)}{\mu c}, \quad \lambda(y) = \frac{k(y)}{\mu c},$$
$$\Gamma(x) = cg\rho_{\text{ref}}(x), \quad \gamma^\epsilon(x) = cg\sigma^\epsilon_{\text{ref}}(x), \quad \gamma(y) = cg\sigma_{\text{ref}}(y).$$

Finally, let χ_f^ϵ denote the characteristic function of Ω_f^ϵ.

Assume throughout that $f \in L^2(J; L^2(\Omega))$, $\rho_{\text{init}} \in H^1(\Omega)$, $\rho_{\text{ref}} \in W^{1,\infty}(\Omega)$, and $\sigma_{\text{ref}} \in W^{1,\infty}(\mathcal{Q}_m)$. The latter two conditions can be relaxed somewhat (see [3]).

THEOREM 1. *For each ϵ, there exists a unique solution to the microscopic model, and $\rho^\epsilon \in H^1(J; L^2(\Omega_f^\epsilon)) \cap L^\infty(J; H^1(\Omega_f^\epsilon))$ and $\sigma^\epsilon \in H^1(J; L^2(\Omega_m^\epsilon)) \cap L^\infty(J; H^1(\Omega_m^\epsilon))$.*

Proof. Recall that ν is the outer unit normal. Then for $\varphi \in H^1(\Omega)$,

(4.1)
$$\begin{aligned}
&-\int_{\partial\Omega_f^\epsilon} \Lambda(x)[\nabla\rho^\epsilon(x,t) + \Gamma(x)(2\rho^\epsilon(x,t) - \rho_{\text{ref}}(x))] \cdot \nu\, \varphi(x)\, ds(x) \\
&\quad = \int_{\partial\Omega_m^\epsilon} \epsilon\lambda^\epsilon(x)[\epsilon\nabla\sigma^\epsilon(x,t) + \gamma^\epsilon(x)(2\sigma^\epsilon(x,t) - \sigma^\epsilon_{\text{ref}}(x))] \cdot \nu\, \varphi(x)\, ds(x) \\
&\quad = \left(\epsilon\lambda^\epsilon[\epsilon\nabla\sigma^\epsilon + \gamma^\epsilon(2\sigma^\epsilon - \sigma^\epsilon_{\text{ref}})], \nabla\varphi\right)_{\Omega_m^\epsilon} + \left(\epsilon\nabla\cdot\lambda^\epsilon[\epsilon\nabla\sigma^\epsilon + \gamma^\epsilon(2\sigma^\epsilon - \sigma^\epsilon_{\text{ref}})], \varphi\right)_{\Omega_m^\epsilon} \\
&\quad = \left(\epsilon\lambda^\epsilon[\epsilon\nabla\sigma^\epsilon + \gamma^\epsilon(2\sigma^\epsilon - \sigma^\epsilon_{\text{ref}})], \nabla\varphi\right)_{\Omega_m^\epsilon} + (\phi^\epsilon\sigma_t^\epsilon, \varphi)_{\Omega_m^\epsilon} - (f, \varphi)_{\Omega_m^\epsilon}.
\end{aligned}$$

Hence, in weak form the microscopic model is

$$(4.2)\quad (\Phi^*\rho_t^\epsilon, \varphi)_{\Omega_f^\epsilon} + \left(\Lambda[\nabla\rho^\epsilon + \Gamma(2\rho^\epsilon - \rho_{\text{ref}})], \nabla\varphi\right)_{\Omega_f^\epsilon} + (\phi^\epsilon\sigma_t^\epsilon, \varphi)_{\Omega_m^\epsilon} + \left(\epsilon\lambda^\epsilon[\epsilon\nabla\sigma^\epsilon + \gamma^\epsilon(2\sigma^\epsilon - \sigma^\epsilon_{\text{ref}})], \nabla\varphi\right)_{\Omega_m^\epsilon} = (f, \varphi)_\Omega, \quad \varphi \in H^1(\Omega),$$

$$(4.3)\quad (\phi^\epsilon\sigma_t^\epsilon, \psi)_{\Omega_m^\epsilon} + \left(\epsilon\lambda^\epsilon[\epsilon\nabla\sigma^\epsilon + \gamma^\epsilon(2\sigma^\epsilon - \sigma^\epsilon_{\text{ref}})], \nabla\psi\right)_{\Omega_m^\epsilon} = (f, \psi)_{\Omega_m^\epsilon}, \quad \psi \in H_0^1(\Omega_m^\epsilon),$$

$$(4.4)\quad \sigma^\epsilon = \rho^\epsilon \quad \text{for } x \in \partial\Omega_m^\epsilon, \quad t > 0.$$

Note that if we let

$$\theta^\epsilon = \begin{cases} \rho^\epsilon & \text{for } x \in \Omega_f^\epsilon, \\ \sigma^\epsilon & \text{for } x \in \Omega_m^\epsilon, \end{cases}$$

then (4.2) is a weak form of

$$\begin{aligned}
&\alpha^\epsilon \theta_t^\epsilon - \nabla \cdot \kappa^\epsilon [\nabla \theta^\epsilon + \beta^\epsilon (2\theta^\epsilon - \theta_{\text{ref}}^\epsilon)] = f \quad \text{for } x \in \Omega, \quad t > 0, \\
&\kappa^\epsilon [\nabla \theta^\epsilon + \beta^\epsilon (2\theta^\epsilon - \theta_{\text{ref}}^\epsilon)] \cdot \nu = 0 \quad \text{for } x \in \partial\Omega, \quad t > 0, \\
&\theta^\epsilon = \rho_{\text{init}} \quad \text{for } x \in \Omega, \quad t = 0,
\end{aligned}$$

where

$$\begin{aligned}
\alpha^\epsilon &= \chi_f^\epsilon \Phi^* + (1 - \chi_f^\epsilon)\phi^\epsilon, \qquad & \kappa^\epsilon &= \chi_f^\epsilon \Lambda + (1 - \chi_f^\epsilon)\epsilon^2 \lambda^\epsilon, \\
\beta^\epsilon &= \chi_f^\epsilon \Gamma + (1 - \chi_f^\epsilon)\epsilon^{-1}\gamma^\epsilon, \qquad & \theta_{\text{ref}}^\epsilon &= \chi_f^\epsilon \rho_{\text{ref}} + (1 - \chi_f^\epsilon)\sigma_{\text{ref}}^\epsilon.
\end{aligned}$$

This is a single well-posed parabolic problem (with discontinuous coefficients). It is known (and easily shown from the a priori estimates of Lemma 1 below) that there exists a unique solution in $H^1(J; L^2(\Omega)) \cap L^\infty(J; H^1(\Omega))$. By restriction, we obtain ρ^ϵ and σ^ϵ as required. □

For each ϵ, we will define a dilation operator "$\sim$" taking measurable functions on Ω_r^ϵ, $r = f, m$, or blank, to measurable functions on $\Omega \times \mathcal{Q}_r$. First let $c^\epsilon(x)$ denote the lattice translation point of the ϵ-cell domain containing x; that is, $c^\epsilon : \Omega \to \epsilon\mathcal{A}'$ is such that $x \in \epsilon\mathcal{Q} + c^\epsilon(x)$. Since the cells are disjoint and fill up space (after dividing up the boundaries of the cells in some nonoverlapping way), c^ϵ is well defined. Then we can define

$$\tilde{\psi}(x, y) = \psi\big(\epsilon y + c^\epsilon(x)\big),$$

where the ϵ (and the r) is implicit. We can now state our main result.

THEOREM 2. *The solution $(\rho^\epsilon, \sigma^\epsilon)$ of the microscopic model converges as $\epsilon \to 0$ to the unique solution (ρ, σ) of the macroscopic model in the following sense:*

$$\begin{aligned}
\chi_f^\epsilon \Phi^* \rho^\epsilon &\longrightarrow \Phi\rho \quad \text{in } H^1(J; L^2(\Omega)) \text{ weakly}, \\
\chi_f^\epsilon \Lambda[\nabla \rho^\epsilon + \Gamma(2\rho^\epsilon - \rho_{\text{ref}})] &\longrightarrow \frac{K}{\mu c}[\nabla \rho + \Gamma(2\rho - \rho_{\text{ref}})] \quad \text{in } L^2(J; L^2(\Omega)) \text{ weakly}, \\
\tilde{\sigma}^\epsilon &\longrightarrow \sigma \quad \text{in } L^2(\Omega; H^1(\mathcal{Q}_m \times J)) \text{ weakly}.
\end{aligned}$$

The proof will be accomplished in several stages. Throughout, C will denote a generic positive constant, not necessarily the same at each occurrence, which is independent of ϵ.

LEMMA 1.

$$\begin{aligned}
&\|\rho^\epsilon\|_{L^\infty(J; H^1(\Omega_f^\epsilon))} + \|\rho_t^\epsilon\|_{L^2(J; L^2(\Omega_f^\epsilon))} \le C, \\
&\|\sigma^\epsilon\|_{L^\infty(J; L^2(\Omega_m^\epsilon))} + \|\sigma_t^\epsilon\|_{L^2(J; L^2(\Omega_m^\epsilon))} \le C, \\
&\|\nabla \sigma^\epsilon\|_{L^\infty(J; L^2(\Omega_m^\epsilon))} \le C\epsilon^{-1}.
\end{aligned}$$

Proof. These are the standard parabolic energy estimates for (4.2); that is, the estimates given by first taking $\varphi = \theta^\epsilon$ and then taking $\varphi = \theta_t^\epsilon$ (which must of course be done on a smooth dense subspace so that the computations can be performed). Note that the domain Ω is fixed, so that C is indeed independent of ϵ. □

The main properties of the dilation operator that we will need are given by Lemma 2.

LEMMA 2. *If $\psi \in L^2(\Omega)$ and $\varphi \in L^2(\Omega)$ (and r is m, f, or blank), then*

$$
\begin{aligned}
(\tilde{\psi}, \tilde{\varphi})_{\Omega \times \mathcal{Q}_r} &= |\mathcal{Q}|(\psi, \varphi)_{\Omega_r^\epsilon}, \\
\nabla_y \tilde{\psi} &= \epsilon \widetilde{\nabla \psi}, \\
\|\tilde{\psi}\|_{L^2(\Omega \times \mathcal{Q}_r)} &= |\mathcal{Q}|^{1/2} \|\psi\|_{L^2(\Omega_r^\epsilon)}, \\
\|\nabla_y \tilde{\psi}\|_{L^2(\Omega \times \mathcal{Q}_r)} &= \epsilon |\mathcal{Q}|^{1/2} \|\nabla \psi\|_{L^2(\Omega_r^\epsilon)}, \\
(\tilde{\psi}, \varphi)_{\Omega \times \mathcal{Q}} &= (\psi, \tilde{\varphi})_{\Omega \times \mathcal{Q}}.
\end{aligned}
$$

Moreover, if ψ is considered to be an element of $L^2(\Omega \times \mathcal{Q}_r)$ that is constant in y, then $\tilde{\psi} \to \psi$ as $\epsilon \to 0$ in $L^2(\Omega \times \mathcal{Q}_r)$ strongly.

Proof. The first two results are simple computations:

$$
\begin{aligned}
(\tilde{\psi}, \tilde{\varphi})_{\Omega \times \mathcal{Q}_r} &= \int_\Omega \int_{\mathcal{Q}_r} \psi\big(\epsilon y + c^\epsilon(x)\big) \varphi\big(\epsilon y + c^\epsilon(x)\big)\, dy\, dx \\
&= \int_\Omega \int_{\epsilon \mathcal{Q}_r + c^\epsilon(x)} \psi(z) \varphi(z) \epsilon^{-3}\, dz\, dx = |\mathcal{Q}|(\psi, \varphi)_{\Omega_r^\epsilon}; \\
\nabla_y \tilde{\psi}(x, y) &= \epsilon \nabla \psi\big(\epsilon y + c^\epsilon(x)\big) = \epsilon \widetilde{\nabla \psi}(x, y).
\end{aligned}
$$

The next two results follow from the first two. The fifth result is another computation:

$$
\begin{aligned}
(\tilde{\psi}, \varphi)_{\Omega \times \mathcal{Q}} &= \int_\Omega \int_{\mathcal{Q}} \psi\big(\epsilon y + c^\epsilon(x)\big) \varphi(x)\, dy\, dx \\
&= \int_\Omega \epsilon^{-3} \int_{\epsilon \mathcal{Q} + c^\epsilon(x)} \psi(z) \varphi(x)\, dz\, dx \\
&= \int_\Omega \psi(z) \left[\epsilon^{-3} \int_{\epsilon \mathcal{Q} + c^\epsilon(z)} \varphi(x)\, dx \right] dz \\
&= \int_\Omega \psi(z) \int_{\mathcal{Q}} \varphi\big(\epsilon y + c^\epsilon(z)\big)\, dy\, dz = (\psi, \tilde{\varphi})_{\Omega \times \mathcal{Q}}.
\end{aligned}
$$

The strong convergence result is clear from the Dominated Convergence Theorem whenever $\psi \in C_0^\infty(\Omega)$, as then ψ is continuous and bounded on Ω and $\lim_{\epsilon \to 0}\big(\epsilon y + c^\epsilon(x)\big) = x$. Since these functions are dense in $L^2(\Omega)$, the result follows from the equivalence of norms. □

Remark 1. A change of variables shows that the last statement of the lemma is an integral form of Lebesgue's theorem on the differentiation of the integral.

COROLLARY 1.

$$
\begin{aligned}
&\|\tilde{\sigma}^\epsilon\|_{L^2(\Omega; H^1(\mathcal{Q}_m \times J))} \le C, \\
&\|\tilde{\rho}^\epsilon\|_{L^2(\Omega; H^1(J; L^2(\mathcal{Q}_f)))} \le C, \\
&\|\nabla_y \tilde{\rho}^\epsilon\|_{L^2(\Omega; L^2(J; L^2(\mathcal{Q}_f)))} \le C\epsilon.
\end{aligned}
$$

Now by Lemma 1 and Corollary 1, we can extract the following weak limits (for a subsequence of the ϵ's):

$$
\begin{aligned}
&\chi_f^\epsilon \Phi^* \rho^\epsilon \longrightarrow \Phi \rho \quad \text{in } H^1(J; L^2(\Omega)) \text{ weakly}, \\
&\chi_f^\epsilon \Lambda[\nabla \rho^\epsilon + \Gamma(2\rho^\epsilon - \rho_{\text{ref}})] \longrightarrow \xi \quad \text{in } L^2(J; L^2(\Omega)) \text{ weakly}, \\
&\tilde{\sigma}^\epsilon \longrightarrow \sigma \quad \text{in } L^2(\Omega; H^1(\mathcal{Q}_m \times J)) \text{ weakly}, \\
&\tilde{\rho}^\epsilon \longrightarrow \tau \quad \text{in } L^2(\Omega; H^1(\mathcal{Q}_f \times J)) \text{ weakly},
\end{aligned}
$$

where $\tau = \tau(x,t)$ only, since $\mathcal{Q}_f$ is connected. In fact, we claim that $\tau = \rho$. Any $\varphi \in C_0^\infty(\Omega \times J)$ is also in $L^2(\Omega; L^2(J; L^2(\mathcal{Q}_f)))$, so

$$\langle \tilde{\rho}^\epsilon, \varphi \rangle_{\Omega \times \mathcal{Q}_f} \longrightarrow \langle \tau, \varphi \rangle_{\Omega \times \mathcal{Q}_f} = |\mathcal{Q}_f| \langle \tau, \varphi \rangle_\Omega.$$

Now Lemma 2 shows that the left-hand side is

$$\langle \tilde{\rho}^\epsilon, \varphi \rangle_{\Omega \times \mathcal{Q}_f} = \langle \widetilde{\chi_f^\epsilon \rho^\epsilon}, \varphi \rangle_{\Omega \times \mathcal{Q}} = \langle \chi_f^\epsilon \rho^\epsilon, \tilde{\varphi} \rangle_{\Omega \times \mathcal{Q}};$$

furthermore,

$$\langle \tilde{\rho}^\epsilon, \varphi \rangle_{\Omega \times \mathcal{Q}_f} = \langle \chi_f^\epsilon \rho^\epsilon, \tilde{\varphi} \rangle_{\Omega \times \mathcal{Q}} \longrightarrow |\mathcal{Q}| \langle (\Phi^*)^{-1} \Phi \rho, \varphi \rangle_\Omega = |\mathcal{Q}_f| \langle \rho, \varphi \rangle_\Omega,$$

so that $\tau = \rho$.

We will now find an equation satisfied by σ. This can be done easily using Lemma 2. Take any $\psi \in L^2(\Omega; L^2(J; H_0^1(\mathcal{Q}_m)))$. Let

$$\hat{\psi}(x,z,t) = \begin{cases} \psi\left(x, \dfrac{z - c^\epsilon(x)}{\epsilon}, t\right) & \text{for } z \in \epsilon \mathcal{Q}_m + c^\epsilon(x), \\ 0 \quad \text{for } z \notin \epsilon \mathcal{Q}_m + c^\epsilon(x). \end{cases}$$

Now for almost every fixed $x \in \Omega$, replace the test function in (4.3) by $\hat{\psi}(x, \cdot, t)$ to see that

$$\begin{aligned} \int_{\epsilon \mathcal{Q}_m + c^\epsilon(x)} & \{ \phi^\epsilon(z) \sigma_t^\epsilon(z,t) \hat{\psi}(x,z,t) \\ & + \epsilon \lambda^\epsilon(z) \big[\epsilon \nabla \sigma^\epsilon(z,t) + \gamma^\epsilon(z) \big(2\sigma^\epsilon(z,t) - \sigma_{\text{ref}}^\epsilon(z) \big) \big] \cdot \nabla_z \hat{\psi}(x,z,t) \} \, dz \\ = & \int_{\epsilon \mathcal{Q}_m + c^\epsilon(x)} f(z,t) \hat{\psi}(x,z,t) \, dz. \end{aligned}$$

Upon dilation (i.e., $z \mapsto \epsilon y + c^\epsilon(x)$) and integration in x and t, we obtain

$$\langle \phi \tilde{\sigma}_t^\epsilon, \psi \rangle_{\Omega \times \mathcal{Q}_m} + \langle \lambda [\nabla_y \tilde{\sigma}^\epsilon + \gamma (2\tilde{\sigma}^\epsilon - \sigma_{\text{ref}})], \nabla_y \psi \rangle_{\Omega \times \mathcal{Q}_m} = \langle \tilde{f}, \psi \rangle_{\Omega \times \mathcal{Q}_m},$$

where ϕ, λ, γ, and σ_{ref} are integrated over $\mathcal{Q}_m$ by their periodicity. As $\epsilon \to 0$, we see that

$$\langle \phi \sigma_t, \psi \rangle_{\Omega \times \mathcal{Q}_m} + \langle \lambda [\nabla_y \sigma + \gamma (2\sigma - \sigma_{\text{ref}})], \nabla_y \psi \rangle_{\Omega \times \mathcal{Q}_m} = \langle f, \psi \rangle_{\Omega \times \mathcal{Q}_m}.$$

This is a weak form of (3.8).

An equation for ρ and ξ can also be derived easily. In (4.2) choose a test function $\varphi \in L^2(J; H^1(\Omega))$ and then integrate in time. The result is that

$$\begin{aligned} & \langle \Phi^* \rho_t^\epsilon, \varphi \rangle_{\Omega_f^\epsilon} + \langle \Lambda [\nabla \rho^\epsilon + \Gamma (2\rho^\epsilon - \rho_{\text{ref}})], \nabla \varphi \rangle_{\Omega_f^\epsilon} \\ & \quad + \langle \phi^\epsilon \sigma_t^\epsilon, \varphi \rangle_{\Omega_m^\epsilon} + \langle \epsilon \lambda^\epsilon [\epsilon \nabla \sigma^\epsilon + \gamma^\epsilon (2\sigma^\epsilon - \sigma_{\text{ref}}^\epsilon)], \nabla \varphi \rangle_{\Omega_m^\epsilon} = \langle f, \varphi \rangle_\Omega. \end{aligned}$$

The first two terms on the left-hand side above tend to $\langle \Phi \rho_t, \varphi \rangle_\Omega + \langle \xi, \nabla \varphi \rangle_\Omega$, whereas the fourth term tends to zero by Lemma 1 since it has an extra power of ϵ. The third term can be dilated to see its convergence by Lemma 2:

$$\langle \phi^\epsilon \sigma_t^\epsilon, \varphi \rangle_{\Omega_m^\epsilon} = |\mathcal{Q}|^{-1} \langle \phi \tilde{\sigma}_t^\epsilon, \tilde{\varphi} \rangle_{\Omega \times \mathcal{Q}_m} \longrightarrow |\mathcal{Q}|^{-1} \langle \phi \sigma_t, \varphi \rangle_{\Omega \times \mathcal{Q}_m} = -\langle f_m, \varphi \rangle_\Omega.$$

Hence, we have shown that

$$\langle \Phi \rho_t, \varphi \rangle_\Omega + \langle \xi, \nabla \varphi \rangle_\Omega = \langle f, \varphi \rangle_\Omega + \langle f_m, \varphi \rangle_\Omega, \qquad \varphi \in L^2(J; H^1(\Omega)). \tag{4.5}$$

Next we will relate ξ to ρ. To this end, let us define $\omega_j^\epsilon \in H^1(\Omega)$ by

$$\omega_j^\epsilon(x) = \epsilon \mathcal{E} \omega_j \left(\frac{x - c^\epsilon(x)}{\epsilon} \right),$$

where $\mathcal{E} : H^1(\mathcal{Q}_f) \to H^1(\mathcal{Q})$ is some bounded extension operator (for example, the one of Calderón [8]). Note that $\tilde{\omega}_j^\epsilon(x,y) = \epsilon \mathcal{E}\omega_j(y)$ and $\widetilde{\nabla\omega_j^\epsilon}(x,y) = \nabla \mathcal{E}\omega_j(y)$. Also let

$$w_{ij} = \frac{1}{|\mathcal{Q}|}\int_{\mathcal{Q}_f} \partial_i \omega_j(y)\, dy.$$

LEMMA 3.

$$\begin{aligned}
&\omega_j^\epsilon \longrightarrow 0 \quad \textit{in } L^2(\Omega) \textit{ strongly},\\
&\epsilon\nabla\omega_j^\epsilon \longrightarrow 0 \quad \textit{in } L^2(\Omega) \textit{ strongly},\\
&\chi_f^\epsilon \partial_i \omega_j^\epsilon \longrightarrow w_{ij} \quad \textit{in } L^2(\Omega) \textit{ weakly}.
\end{aligned}$$

Proof. For the first limit, note that

$$\|\omega_j^\epsilon\|_{L^2(\Omega)} = |\mathcal{Q}|^{-1/2}\|\tilde{\omega}_j^\epsilon\|_{L^2(\Omega\times\mathcal{Q})} = |\mathcal{Q}|^{-1/2}|\Omega|\,\epsilon\,\|\mathcal{E}\omega_j\|_{L^2(\mathcal{Q})} \le C\epsilon\|\omega_j\|_{L^2(\mathcal{Q}_f)} \longrightarrow 0.$$

The second limit is similar, noting that

$$\|\nabla\omega_j^\epsilon\|_{L^2(\Omega)} = |\mathcal{Q}|^{-1/2}\|\widetilde{\nabla\omega_j^\epsilon}\|_{L^2(\Omega\times\mathcal{Q})} = |\mathcal{Q}|^{-1/2}|\Omega|\,\|\nabla\mathcal{E}\omega_j\|_{L^2(\mathcal{Q})} \le C\|\nabla\omega_j\|_{L^2(\mathcal{Q}_f)}.$$

The above expression bounds the $L^2(\Omega)$-norm of $\chi_f^\epsilon\partial_i\omega_j^\epsilon$, so this expression has some weak limit. To see what this limit is, solve the elliptic problem modulo constants

$$\begin{aligned}
&-\Delta\psi = \chi_{\mathcal{Q}_f}\partial_i\omega_j - w_{ij} \quad \text{for } y \in \mathcal{Q},\\
&\nabla\psi\cdot\nu = 0 \quad \text{for } y \in \partial\mathcal{Q},
\end{aligned}$$

where $\chi_{\mathcal{Q}_f}$ is the characteristic function of $\mathcal{Q}_f$. We can do this since the average of $\chi_{\mathcal{Q}_f}\partial_i\omega_j - w_{ij}$ is zero. Note that $\psi \in H^1(\mathcal{Q})$. Now for $\varphi \in C_0^\infty$,

$$\begin{aligned}
(\chi_f^\epsilon\partial_i\omega_j^\epsilon - w_{ij},\varphi)_\Omega &= |\mathcal{Q}|^{-1}(\tilde{\chi}_f^\epsilon\widetilde{\partial_i\omega_j^\epsilon} - w_{ij},\tilde{\varphi})_{\Omega\times\mathcal{Q}} = |\mathcal{Q}|^{-1}(\chi_{\mathcal{Q}_f}\partial_i\omega_j - w_{ij},\tilde{\varphi})_{\Omega\times\mathcal{Q}}\\
&= |\mathcal{Q}|^{-1}(-\Delta\psi,\tilde{\varphi})_{\Omega\times\mathcal{Q}} = |\mathcal{Q}|^{-1}(\nabla\psi,\epsilon\widetilde{\nabla\varphi})_{\Omega\times\mathcal{Q}} \le C\epsilon \longrightarrow 0. \qquad \square
\end{aligned}$$

From the definition (3.2)–(3.3) of ω_j we can see that

$$0 = -\big(\nabla\cdot(\nabla\omega_j^\epsilon + e_j),\psi\big)_{\Omega_f^\epsilon} = (\nabla\omega_j^\epsilon + e_j,\nabla\psi)_{\Omega_f^\epsilon}, \qquad \psi \in H_0^1(\Omega). \tag{4.6}$$

For any $\varphi \in C_0^\infty$, take $\psi = \rho^\epsilon\Lambda\varphi$ above. After adding and subtracting the same term twice and after integrating in time, we have that

$$\begin{aligned}
&\langle\nabla\omega_j^\epsilon,\rho^\epsilon\nabla(\Lambda\varphi) - \varphi\Lambda\Gamma(2\rho^\epsilon - \rho_{\text{ref}})\rangle_{\Omega_f^\epsilon} + \langle e_j,\rho^\epsilon\nabla(\Lambda\varphi) - \varphi\Lambda\Gamma(2\rho^\epsilon - \rho_{\text{ref}})\rangle_{\Omega_f^\epsilon}\\
&\quad + \langle\nabla\omega_j^\epsilon,\varphi\Lambda[\nabla\rho^\epsilon + \Gamma(2\rho^\epsilon - \rho_{\text{ref}})]\rangle_{\Omega_f^\epsilon} + \langle e_j,\varphi\Lambda[\nabla\rho^\epsilon + \Gamma(2\rho^\epsilon - \rho_{\text{ref}})]\rangle_{\Omega_f^\epsilon} = 0.
\end{aligned} \tag{4.7}$$

The first term converges by dilation:

$$\begin{aligned}
&\langle\nabla\omega_j^\epsilon,\rho^\epsilon\nabla(\Lambda\varphi) - \varphi\Lambda\Gamma(2\rho^\epsilon - \rho_{\text{ref}})\rangle_{\Omega_f^\epsilon}\\
&\quad = |\mathcal{Q}|^{-1}\langle\nabla\omega_j,\tilde{\rho}^\epsilon\widetilde{\nabla(\Lambda\varphi)} - \widetilde{\varphi\Lambda}\Gamma(2\tilde{\rho}^\epsilon - \tilde{\rho}_{\text{ref}})\rangle_{\Omega\times\mathcal{Q}_f}\\
&\quad \longrightarrow |\mathcal{Q}|^{-1}\langle\nabla\omega_j,\rho\nabla(\Lambda\varphi) - \varphi\Lambda\Gamma(2\rho - \rho_{\text{ref}})\rangle_{\Omega\times\mathcal{Q}_f}\\
&\quad = \sum_{i=1}^{3}\langle w_{ij},\rho\partial_i(\Lambda\varphi) - \varphi\Lambda\Gamma_i(2\rho - \rho_{\text{ref}})\rangle_\Omega.
\end{aligned}$$

The second term converges trivially to

$$\frac{|\mathcal{Q}_f|}{|\mathcal{Q}|}\langle e_j,\rho\nabla(\Lambda\varphi) - \varphi\Lambda\Gamma(2\rho - \rho_{\text{ref}})\rangle_\Omega$$

since $\chi_f^\epsilon \rightharpoonup |\mathcal{Q}_f|/|\mathcal{Q}|$ in $L^2(\Omega)$ weakly is easily shown from Lemma 2. The convergence of the third term is found from (4.2) by taking the test function $\omega_j^\epsilon \varphi$. Integration in time shows that

$$\langle \Phi^* \rho_i^\epsilon, \omega_j^\epsilon \varphi \rangle_{\Omega_f^\epsilon} + \langle \Lambda[\nabla \rho^\epsilon + \Gamma(2\rho^\epsilon - \rho_{\text{ref}})], \omega_j^\epsilon \nabla \varphi + \varphi \nabla \omega_j^\epsilon \rangle_{\Omega_f^\epsilon}$$
$$+ \langle \phi^\epsilon \sigma_i^\epsilon, \omega_j^\epsilon \varphi \rangle_{\Omega_m^\epsilon} + \langle \epsilon \lambda^\epsilon [\epsilon \nabla \sigma^\epsilon + \gamma^\epsilon (2\sigma^\epsilon - \sigma_{\text{ref}}^\epsilon)], \omega_j^\epsilon \nabla \varphi + \varphi \nabla \omega_j^\epsilon \rangle_{\Omega_m^\epsilon} = \langle f, \omega_j^\epsilon \varphi \rangle_\Omega.$$

Lemmas 1 and 3 now show that

$$\lim_{\epsilon \to 0} \langle \nabla \omega_j^\epsilon, \varphi \Lambda [\nabla \rho^\epsilon + \Gamma(2\rho^\epsilon - \rho_{\text{ref}})] \rangle_{\Omega_f^\epsilon} = 0.$$

Hence we conclude that the fourth term of (4.7) converges to

$$\langle \xi_j, \varphi \rangle_\Omega$$
$$= \sum_{i=1}^{3} \left\{ - \left\langle \left[\frac{|\mathcal{Q}_f|}{|\mathcal{Q}|} \delta_{ij} + w_{ij} \right] \rho, \partial_i(\Lambda \varphi) \right\rangle_\Omega + \left\langle \left[\frac{|\mathcal{Q}_f|}{|\mathcal{Q}|} \delta_{ij} + w_{ij} \right] \Lambda \Gamma_i (2\rho - \rho_{\text{ref}}), \varphi \right\rangle_\Omega \right\},$$

which is the distributional form of

$$\xi = \frac{K}{\mu c} [\nabla \rho + \Gamma(2\rho - \rho_{\text{ref}})].$$

Hence (4.5) is a weak form of (3.5)–(3.6), (3.11).

The following trivial proposition enables us to derive the boundary and initial conditions satisfied by our limit functions.

PROPOSITION. *If $\mathcal{T} : X \to Y$ is a continuous linear operator between Banach spaces X and Y, and if $\psi^\epsilon \rightharpoonup \psi$ in X weakly, then $\mathcal{T}\psi^\epsilon \rightharpoonup \mathcal{T}\psi$ in Y weakly.*

If $\mathcal{T}_0$ is the linear operator giving the trace for time zero, then [15] $\mathcal{T}_0 : H^1(\Omega \times J) \to H^{1/2}(\Omega)$ is bounded. Hence

$$\mathcal{T}_0(\chi_f^\epsilon \Phi^* \rho^\epsilon) \longrightarrow \mathcal{T}_0(\Phi \rho) = \Phi \mathcal{T}_0 \rho.$$

But

$$\mathcal{T}_0(\chi_f^\epsilon \Phi^* \rho^\epsilon) = \chi_f^\epsilon \Phi^* \rho_{\text{init}} \longrightarrow \Phi \rho_{\text{init}};$$

consequently, we have (3.7). For almost every x, we also have (3.10) since, for the appropriate trace operator,

$$\mathcal{T}_0 \tilde{\sigma}^\epsilon(x, y) \longrightarrow \mathcal{T}_0 \sigma(x, y)$$

and

$$\mathcal{T}_0 \tilde{\sigma}^\epsilon(x, y) = \tilde{\rho}_{\text{init}}(x, y) \longrightarrow \rho_{\text{init}}(x).$$

If now $\mathcal{T}_b : H^1(\mathcal{Q}_r \times J) \to H^{1/2}(\partial \mathcal{Q}_r \times J)$, $r = m$ or f, is the boundary trace operator, then for almost every x

$$\mathcal{T}_b \tilde{\sigma}^\epsilon(x, y, t) \longrightarrow \mathcal{T}_b \sigma(x, y, t)$$

and

$$\mathcal{T}_b \tilde{\sigma}^\epsilon(x, y, t) = \mathcal{T}_b \tilde{\rho}^\epsilon(x, y, t) \longrightarrow \mathcal{T}_b \rho(x, t) = \rho(x, t).$$

This shows (3.9).

We now have a solution to the macroscopic model. To see that this solution is unique, assume that ρ and σ are the differences of two solutions. These then satisfy

the macroscopic model's equations with $f \equiv \rho_{\text{init}} \equiv 0$ and the gravitational pseudo-source terms $\nabla \cdot \left((K/\mu) g \rho_{\text{ref}}^2\right)$ and $\nabla \cdot \left((k/\mu) g \sigma_{\text{ref}}^2\right)$ set to zero. Multiply (3.5) by ρ and integrate in x to see that

$$(\Phi \rho_t, \rho)_\Omega + \left(\frac{K}{\mu c}\left[\nabla \rho + 2cg\rho\rho_{\text{ref}}\right], \nabla \rho\right)_\Omega = (f_m, \rho)_\Omega. \tag{4.8}$$

Now multiply (3.8) by $\sigma - \rho$ and integrate in both x and y to obtain

$$(\phi\sigma_t, \sigma)_{\Omega \times \mathcal{Q}_m} - (\phi\sigma_t, \rho)_{\Omega \times \mathcal{Q}_m} + \left(\frac{k}{\mu c}\left[\nabla_y \sigma + 2cg\sigma\sigma_{\text{ref}}\right], \nabla_y \sigma\right)_{\Omega \times \mathcal{Q}_m} = 0, \tag{4.9}$$

since $\nabla_y \rho = 0$. Note that $f_m = -|\mathcal{Q}|^{-1}(\phi\sigma_t, 1)_{\mathcal{Q}_m}$ cancels when $|\mathcal{Q}|$ times the first equation is added to the second. Hence standard energy estimates of this combined equation show the uniqueness. Finally, since the solution is unique, the entire sequence of solutions to the microscopic model converges as required, and the proof of Theorem 2 is complete.

Of course, our macroscopic coefficients should have the appropriate properties. Obviously Φ is uniformly positive. As for K, we have Theorem 3.

THEOREM 3. *The macroscopic fracture permeability tensor $K(x)$ is symmetric and positive definite.*

Proof. From (3.2)–(3.3), we see that

$$0 = -\left(\nabla \cdot (\nabla \omega_j + e_j), \omega_i\right)_{\mathcal{Q}_f} = (\nabla \omega_j + e_j, \nabla \omega_i)_{\mathcal{Q}_f}$$

by the periodicity of the ω_k. Hence

$$(\partial_j \omega_i, 1)_{\mathcal{Q}_f} = -(\nabla \omega_j, \nabla \omega_i)_{\mathcal{Q}_f},$$

which shows that K is symmetric. In fact, we can use this to rewrite K_{ij}:

$$\begin{aligned} |\mathcal{Q}|(K^*)^{-1}K_{ij} &= |\mathcal{Q}_f|\delta_{ij} + (\partial_i \omega_j, 1)_{\mathcal{Q}_f} \\ &= (\nabla y_j, \nabla y_i)_{\mathcal{Q}_f} + (\partial_i \omega_j, 1)_{\mathcal{Q}_f} + (\partial_j \omega_i, 1)_{\mathcal{Q}_f} - (\partial_j \omega_i, 1)_{\mathcal{Q}_f} \\ &= (\nabla y_j, \nabla y_i)_{\mathcal{Q}_f} + (\nabla \omega_j, \nabla y_i)_{\mathcal{Q}_f} + (\nabla \omega_i, \nabla y_j)_{\mathcal{Q}_f} + (\nabla \omega_j, \nabla \omega_i)_{\mathcal{Q}_f} \\ &= \left(\nabla(y_j + \omega_j), \nabla(y_i + \omega_i)\right)_{\mathcal{Q}_f}. \end{aligned}$$

This shows that K is positive semidefinite. Definiteness follows from the connectedness of $\mathcal{Q}_f$ and the periodicity of the ω_j. Let ξ be any constant vector. Then, with ω being the vector whose components are the ω_j,

$$0 = \sum_{i,j} \xi_j \left(\nabla(y_j + \omega_j), \nabla(y_i + \omega_i)\right)_{\mathcal{Q}_f} \xi_i = \sum_k \left(\partial_k[\xi \cdot (y + \omega)], \partial_k[\xi \cdot (y + \omega)]\right)_{\mathcal{Q}_f}.$$

Since $\mathcal{Q}_f$ is connected, $\omega(y) \cdot \xi = \alpha - y \cdot \xi$ for some constant α. This being periodic in y forces ξ to be zero. □

Remark 2. The tensor K may not be strictly positive definite for some degenerate (disconnected) geometries. For example, in one space dimension $\mathcal{Q}_f$ consists of two disjoint intervals, so $K \equiv 0$.

We close by noting that the macroscopic model is well posed.

THEOREM 4. *If $\partial\Omega$ is smooth, then $\rho \in H^1(J; L^2(\Omega)) \cap L^2(J; H^2(\Omega))$ and $\sigma \in L^2(\Omega; H^1(J; L^2(\mathcal{Q}_m))) \cap L^2(\Omega; L^2(J; H^2(\mathcal{Q}_m)))$. Moreover, the solution varies continuously with the data $f \in L^2(J; L^2(\Omega))$, $\rho_{\text{init}} \in H^1(\Omega)$, $\rho_{\text{ref}} \in W^{1,\infty}(\Omega)$, and $\sigma_{\text{ref}} \in W^{1,\infty}(\mathcal{Q}_m)$.*

Proof. The theorem was proven by Arbogast for the original formulation of the model using the method of continuity [3]. He also proved it for a generalization of the

model that included gravity in the matrix (and had a more complicated definition of the matrix source term in that the fracture density was assumed to vary linearly in space over each block) [2]. This latter proof is the most convenient for the present situation. It treats the matrix source term in much the same way as we did in proving uniqueness of the solution; that is, it drops out of the equations after properly combining (3.5) and (3.8). The external source in the matrix equations can be dealt with by combining it with the gravitational pseudo-source term. By properly taking into account the subtle differences between the homogenized and original formulations of the model mentioned at the end of §3 (and by noting that the fracture flow is simply constant in space over each block), the proof goes through easily. □

REFERENCES

[1] B. AMAZIANE AND A. BOURGEAT, *Effective behavior of two-phase flow in heterogeneous reservoir*, Numerical Simulation in Oil Recovery, M. F. Wheeler, ed., IMA Volumes in Mathematics and Its Applications 11, Springer-Verlag, Berlin, New York, 1988, pp. 1–22.

[2] T. ARBOGAST, *The double porosity model for single phase flow in naturally fractured reservoirs*, Numerical Simulation in Oil Recovery, M. F. Wheeler, ed., the IMA Volumes in Mathematics and Its Applications 11, Springer-Verlag, Berlin, New York, 1988, pp. 23–45.

[3] ———, *Analysis of the simulation of single phase flow through a naturally fractured reservoir*, SIAM J. Numer. Anal., 26 (1989), pp. 12–29.

[4] T. ARBOGAST, J. DOUGLAS, JR., AND U. HORNUNG, *Modeling of naturally fractured reservoirs by formal homogenization techniques*, to appear.

[5] G. I. BARENBLATT, IU. P. ZHELTOV, AND I. N. KOCHINA, *Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks [strata]*, Prikl. Mat. Mekh., 24 (1960), pp. 852–864. (In Russian.) J. Appl. Math. Mech., 24 (1960), pp. 1286–1303.

[6] A. BENSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[7] A. BOURGEAT, *Homogenization of two phase flow equations*, Proc. Sympos. Pure Math., 45 (1986), pp. 157–163.

[8] A. P. CALDERÓN, *Lebesgue spaces of differentiable functions and distributions*, Proc. Sympos. Pure Math., 4 (1961), pp. 33–49.

[9] J. DOUGLAS, JR., AND T. ARBOGAST, *Dual-porosity models for flow in naturally fractured reservoirs*, in Dynamics of Fluids in Hierarchical Porous Media, J. H. Cushman, ed., Academic Press, London, 1990, pp. 177–221.

[10] J. DOUGLAS, JR., P. J. PAES LEME, T. ARBOGAST, AND T. SCHMITT, *Simulation of flow in naturally fractured reservoirs*, Paper SPE 16019, Proc., 9th SPE Symposium on Reservoir Simulation, Society of Petroleum Engineers, Dallas, TX, 1987, pp. 271–279.

[11] U. HORNUNG, *Miscible displacement in porous media influenced by mobile and immobile water*, Tech. Report 102, Department of Mathematics, Arizona State University, Tempe, AZ, 1987.

[12] U. HORNUNG AND W. JÄGER, *A model for chemical reactions in porous media*, in Complex Chemical Reaction Systems. Mathematical Modelling and Simulation, J. Warnatz and W. Jäger, eds., Springer Series in Chemical Physics 47, Springer-Verlag, Berlin, New York, 1987, pp. 318–334.

[13] U. HORNUNG AND R. SHOWALTER, *Diffusion models for fractured media*, J. Math. Anal. Appl., to appear.

[14] H. KAZEMI, *Pressure transient analysis of naturally fractured reservoirs with uniform fracture distribution*, Soc. Petroleum Engrs. J., 9 (1969), pp. 451–462.

[15] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applcations* 1, Springer-Verlag, Berlin, New York, 1970.

[16] S. J. PIRSON, *Performance of fractured oil reservoirs*, Bull. Amer. Assoc. Petroleum Geologists, 37 (1953), pp. 232–244.

[17] E. SANCHEZ-PALENCIA, *Non-homogeneous Media and Vibration Theory*, Lecture Notes in Physics 127, Springer-Verlag, Berlin, New York, 1980.

[18] A. DE SWAAN, *Analytic solutions for determining naturally fractured reservoir properties by well testing*, Soc. Petroleum Engrs. J., 16 (1976), pp. 117–122.

[19] L. TARTAR, *Incompressible fluid flow in a porous medium–convergence of the homogenization process*, in Non-homogeneous Media and Vibration Theory, E. Sanchez-Palencia, Lecture Notes in Physics 127, Springer-Verlag, Berlin, New York, 1980, pp. 368–377.

[20] CH. VOGT, *A homogenization theorem leading to a Volterra integro-differential equation for permeation chromatography*, Preprint 155, Sonderforschungsbereich 123, Heidelberg, FRG, 1982.

[21] J. E. WARREN AND P. J. ROOT, *The behavior of naturally fractured reservoirs*, Soc. Petroleum Engrs. J., 3 (1963), pp. 245–255.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 837-866, July 1990
© 1990 Society for Industrial and Applied Mathematics
002

TRANSITIONAL WAVES FOR CONSERVATION LAWS*

ELI L. ISAACSON†, DAN MARCHESIN‡, AND BRADLEY J. PLOHR§

Abstract. A new class of fundamental waves arises in conservation laws that are not strictly hyperbolic. These waves serve as transitions between wave groups associated with particular characteristic families. Transitional shock waves are discontinuous solutions that possess viscous profiles but do not conform to the Lax characteristic criterion; they are sensitive to the precise form of the physical viscosity. Transitional rarefaction waves are rarefaction fans across which the characteristic family changes from faster to slower.

This paper identifies an extensive family of transitional shock waves for conservation laws with quadratic fluxes and arbitrary viscosity matrices; this family comprises all transitional shock waves for a certain class of such quadratic models. The paper also establishes, for general systems of two conservation laws, the generic nature of rarefaction curves near an elliptic region, thereby identifying transitional rarefaction waves. The use of transitional waves in solving Riemann problems is illustrated by an example where the characteristic and viscous profile admissibility criteria yield distinct solutions.

Key words. conservation laws, Riemann problems, nonstrictly-hyperbolic, admissibility, viscosity, saddle-saddle connections, quadratic dynamical systems, line fields

AMS(MOS) subject classifications. 34D30, 35L65, 35L67, 35L80, 58F09

1. Introduction.

Non-strictly hyperbolic systems of conservation laws possess fundamental wave solutions that are distinct from classical rarefaction and shock waves. These new waves are not associated with a particular characteristic family; rather, they serve as transitions between classical wave groups. In the presence of such transitional waves, the solution of a Riemann problem for a system of n conservation laws can contain more than n wave groups. The purpose of the present paper is to study the character of transitional waves and the crucial role they play in solving Riemann problems.

For a particular model system of two conservation laws, Shearer et al. [35] found it impossible to solve the general Riemann problem using only classical Lax shock waves. However, the general solution exists and is unique provided that a limited family of nonclassical discontinuities is allowed. For these crossing discontinuities, neither family of characteristics is compressive, in contrast to shock waves. The same type of discontinuity occurs in reactive gas dynamics as weak deflagration waves [5]. To solve another model system, Isaacson and Temple [21] use rarefaction waves that switch from characteristic family 2 to family 1 at the locus where eigenvalues coincide. Composite waves built from such discontinuities and rarefaction waves also arise [17], [14]. Thus the solution of a Riemann problem can involve waves that are not associated with a unique characteristic family. We view these examples as instances of a new class of waves, transitional waves. Thus a transitional shock wave is a crossing discontinuity that conforms to an admissibility criterion, and a transitional rarefaction wave changes from a faster family to a slower family. The nature of transitional shock and rarefaction waves is the subject of this paper.

* Received by the editors November 2, 1988; accepted for publication (in revised form) June 11, 1989. This work was supported by the National Science Foundation, the Army Research Office, the Financiadora de Estudos e Pesquisas, and the Conselho Nacional de Pesquisas.

† Department of Mathematics, P.O. Box 3036 University Station, University of Wyoming, Laramie, Wyoming 82071.

‡ Instituto de Matemática Pura e Aplicada and Pontifícia Universidade Católica, Rio de Janeiro, 22460 RJ, Brazil.

§ Computer Sciences Department and Center for the Mathematical Sciences, University of Wisconsin, Madison, Wisconsin 53706.

As discussed in § 2, we allow for any discontinuity that is the limit of traveling wave solutions of the conservation laws as augmented by a parabolic term; this is motivated by physical considerations. Admissible discontinuities correspond, then, to orbits connecting singularities of a dynamical system associated with the parabolic system. While a classical shock wave generalizes to an orbit between a node and a saddlepoint, which is structurally stable under general perturbations, an admissible crossing discontinuity corresponds to a saddle-saddle connection, which has stringent stability restrictions. In particular, the class of transitional shock waves depends critically on the viscous term in the parabolic equation. Furthermore, saddle-saddle connections signal bifurcations of admissible discontinuities. Also possible are waves corresponding to connecting orbits between nodes; however, these totally compressive waves can be decomposed into classical shock waves.

In § 3 we study transitional shock waves for conservation laws with quadratic fluxes, which are simple models in which the new shock waves occur. The analysis is based on an explicit calculation that establishes a direct relationship between transitional shock waves and the viscous term: for a class of such quadratic models, a saddle-saddle connection must lie along a straight line parallel to a direction associated with the viscosity matrix. We provide a complete characterization of transitional shock waves of this form in §§ 3.1–3.3; circumstances under which these are the only transitional shock waves are established in § 3.3 and Appendix A. As a consequence, the regions in state space where these waves play a role are easily identified. The case where the viscosity matrix is the identity, which is commonly assumed in analyses of traveling waves, is shown to have degenerate features.

In § 4 we determine the behavior of rarefaction curves near a boundary between elliptic and hyperbolic behavior in a general system of two conservation laws. The transitional rarefaction waves are shown to arise from integral curves through isolated points on this boundary. The method of analysis represents both the 1- and 2-family rarefaction curves as a foliation defined by a single line field on a certain manifold. For strictly hyperbolic conservation laws, this manifold consists of two separate sheets, one for each family; in mixed-type problems, the two sheets are joined at the elliptic-hyperbolic boundary. Examples of these manifolds are given in Appendix B.

We illustrate the use of transitional waves to solve Riemann problems in § 5. The essential construct is the transitional curve. The examples we study belong to a class of quadratic models for which the Riemann problem has a unique solution using only classical shock waves [21], [32]. We show, however, that some of the classical shock waves used in these solutions do not admit viscous profiles, and that transitional shock waves can be used in their place.

Finally, in § 6, we summarize our results.

2. Transitional waves. In this section we study the structure of solutions of Riemann problems. If the states in the initial data are close, the solution constructed classically consists of several groups of waves, each group corresponding to a characteristic family. Globally, however, the solution might contain transitional waves that interpolate between families, as we describe in § 2.1. To determine an appropriate class of transitional waves, we invoke the admissibility criterion based on viscous profiles; this criterion is discussed in § 2.2.

2.1. Wave groups. We are interested in solutions of a system of conservation laws

$$U_t + F(U)_x = 0 \tag{2.1}$$

governing the evolution, in one space dimension, of an n-dimensional state vector U.

The function F is called the flux. The characteristic speeds for (2.1), i.e., the eigenvalues of the Jacobian derivative matrix $F'(U)$, are denoted $\lambda_i(U)$, $i=1,\dots,n$. In the hyperbolic region, where the characteristic speeds are real, we order them as

$$\lambda_1(U)\leqq\lambda_2(U)\leqq\cdots\leqq\lambda_n(U). \tag{2.2}$$

The dependence of the characteristic speeds on U leads, in general, to focusing of waves and formation of discontinuous solutions, so that (2.1) must be interpreted in the sense of distributions.

Much of the structure of general solutions of (2.1) is reflected in solutions that respect the invariance of the equation under the scaling transformation $(t, x)\mapsto(\alpha t, \alpha x)$. Such scale-invariant solutions satisfy the initial conditions of a Riemann problem: at $t=0$, the solution U must be a constant U_L for $x<0$ and another constant U_R for $x>0$. Conversely, solutions of a Riemann problem are expected to be scale-invariant, i.e., they depend on t and x only through the combination $\xi=x/t$. Although Riemann problems are only special initial value problems, the solutions of the general Cauchy initial value problem may be viewed as a nonlinear superposition of scale-invariant solutions [8].

A scale-invariant solution can be partitioned into several groups of waves; the waves in each group move together as a single entity. More precisely, we define a *wave group* to be a scale-invariant solution that contains no intermediate constant states. Thus a solution of a Riemann problem comprises a sequence of wave groups moving apart from each other, as in Fig. 1a. Wave groups are composed of two basic ingredients: centered rarefaction waves and centered discontinuous waves (see Fig. 1b). A centered rarefaction wave associated with a characteristic family i is constructed using integral curves of the differential equation

$$\dot{U}=r_i(U), \tag{2.3}$$

where $r_i(U)$ is a right eigenvector corresponding to $\lambda_i(U)$. A rarefaction wave corresponds to a segment of an integral curve along which $\lambda_i(U)$ is nondecreasing; it is defined by inverting the relation $\lambda_i(U)=\xi$. A centered discontinuous wave is a jump discontinuity that propagates at speed s and separates two constant states U_- and U_+, where U_-, U_+, and s satisfy the system of n equations

$$-s[U_+-U_-]+F(U_+)-F(U_-)=0, \tag{2.4}$$

called the Rankine-Hugoniot jump condition. (By convention, U_- is on the left side of the discontinuity and U_+ is on the right side.)

The class of allowable discontinuous waves must be restricted according to criteria that reflect physical dissipation mechanisms neglected in the governing equations. For

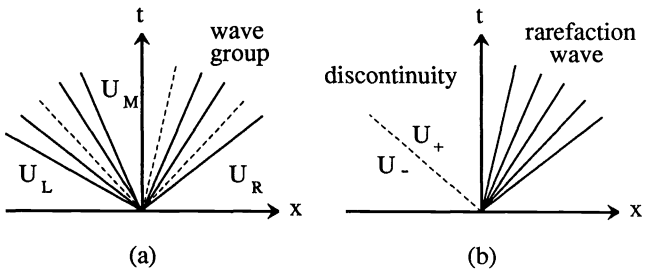


FIG. 1. *Scale-invariant solutions*: (a) *a solution of a Riemann problem, comprising a sequence of wave groups*; (b) *a centered rarefaction wave and a centered discontinuous wave.*

conservation laws that are genuinely nonlinear, Lax [25] has introduced the admissibility requirement that the characteristics of one family impinge on both sides of the discontinuity, while the characteristics of the other families cross through the discontinuity. For more general conservation laws, characteristics must be permitted to become tangent to the discontinuity. Therefore we define a centered discontinuous wave to be a *Lax discontinuity* of the ith family provided that the characteristic speeds are related to the propagation speed as follows:

$$\lambda_i(U_+) \leqq s \leqq \lambda_i(U_-), \tag{2.5}$$

$$\lambda_{i-1}(U_-) \leqq s \leqq \lambda_{i+1}(U_+). \tag{2.6}$$

Remark. A Lax discontinuity is associated with a unique family except in one case: an ith-family discontinuity for which $\lambda_i(U_-) = s = \lambda_{i+1}(U_+)$ may be regarded also as associated with family $i+1$. This ambiguity in nomenclature, however, does not affect our results.

If we adopt the admissibility criterion based on characteristics and assume that all characteristic speeds are distinct, then any wave, i.e., rarefaction wave or discontinuity, has an associated family. Observe that (i) no wave can be preceded by a wave of a faster family; and (ii) two waves of the same family must belong to the same wave group. Therefore a solution of a Riemann problem can contain at most n wave groups. In particular, no wave can appear strictly between a wave group containing an i-wave and another group containing an $(i+1)$-wave. These facts [26] generalize the classical picture [25] in which a solution of a Riemann problem consists of at most n shock or rarefaction waves, separated by constant states, where each wave is associated with a distinct family.

The characteristic criterion, however, is sometimes overly restrictive and at other times too lax—a Riemann problem might have no solution or it might have many. An alternative admissibility criterion is to require discontinuous waves to possess viscous profiles, as described more fully in § 2.2. This is the viscosity admissibility criterion. In general, it is distinct from the characteristic criterion, since there exist Lax discontinuities that do not have viscous profiles, while some discontinuities with viscous profiles are not of Lax type. The viscosity criterion, too, can fail to guarantee existence and uniqueness of solutions of Riemann problems, but we prefer it to the characteristic criterion because it derives from certain physical effects that have been neglected in the modeling equations. In this paper we adopt the viscosity admissibility criterion.

When discontinuities that are not of Lax type are allowed, or when characteristic speeds may coincide, it is possible for the solution of a Riemann problem to contain more than n wave groups. Consider, for example, a discontinuity satisfying

$$\lambda_i(U_-) < s < \lambda_{i+1}(U_-), \tag{2.7}$$

$$\lambda_i(U_+) < s < \lambda_{i+1}(U_+), \tag{2.8}$$

through which all characteristics cross. Such a *crossing discontinuity* can have a viscous profile, as we show in § 3.3. This wave can appear strictly between a wave group containing an i-wave and one containing an $(i+1)$-wave (see Fig. 2a). As another possibility, an integral curve of family $i+1$ might pass tangent to the locus where $\lambda_{i+1}(U) = \lambda_i(U)$ and continue with an integral curve of family i (see § 4). This gives a rarefaction wave that can lie strictly between i-waves and $(i+1)$-waves, as in Fig. 2b.

In general, therefore, a distinct wave group can appear between one group containing an i-wave and another containing an $(i+1)$-wave. We call it an *i, $(i+1)$-transitional wave group.* A *transitional shock wave* is a crossing discontinuity that

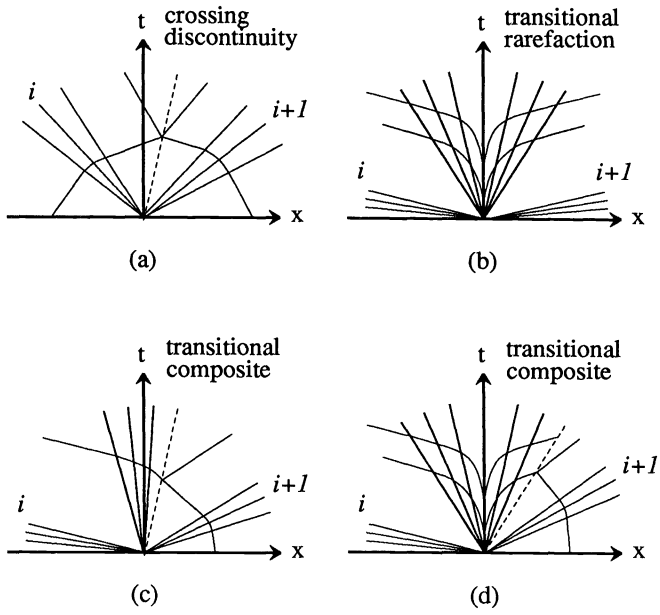


FIG. 2. *Transitional waves*: (a) *a crossing discontinuity*; (b) *a transitional rarefaction wave*; (c), (d) *transitional composite waves. Light lines are characteristics.*

conforms to the viscosity admissibility criterion, and a *transitional rarefaction wave* is a rarefaction wave that changes from a faster family to a slower family. More complicated transitional wave groups are also possible. For example, a discontinuity satisfying $\lambda_i(U_-) < s = \lambda_{i+1}(U_-)$, as well as (2.8), can adjoin an $(i+1)$-wave group on its left [17], as in Fig. 2c. Similarly, a transitional rarefaction can adjoin an i-wave group on tis right (see Fig. 2d) or an $(i+1)$-wave group on its left [14].

2.2. Viscosity admissibility criterion. Typically, (2.1) is an approximation to an equation of the form

$$U_t + F(U)_x = \varepsilon[D(U)U_x]_x \tag{2.9}$$

in the (singular) limit as $\varepsilon \to 0^+$. Here $D(U)$ is the viscosity matrix, which models certain physical effects that are neglected in the conservation law. We require that the eigenvalues of $D(U)$ have positive real part; this guarantees that short-wavelength perturbations of constant solutions decay exponentially in time.

Physically realizable solutions of (2.1) are expected to be limits of solutions of the parabolic equation (2.9). In particular, certain centered discontinuous waves arise as limits of traveling wave solutions in the following way. A traveling wave depends on t and x only through the combination $\xi = (x - st)/\varepsilon$, and it approaches limits U_+ and U_- as $\xi \to \pm\infty$. Therefore (2.9) can be integrated once to obtain the dynamical system

$$-s[U(\xi) - U_-] + F(U(\xi)) - F(U_-) = D(U(\xi))\dot{U}(\xi), \tag{2.10}$$

where the dot denotes differentiation with respect to ξ. Taking the limit of (2.10) as $\xi \to \infty$ shows that U_+, U_-, and s must be related by the Rankine-Hugoniot condition (2.4), so that U_+ and U_- are critical points for the dynamical system. As $\varepsilon \to 0^+$, the spatial region over which the solution makes the transition from U_- to U_+ shrinks to a point at $x = st$. Consequently, the traveling wave solution approaches a centered discontinuous wave. Thus a discontinuity is said to have a *viscous profile* when the

dynamical system (2.10) has a connecting orbit flowing from U_- to U_+. It is natural to regard a discontinuity as admissible only if it has a viscous profile; this is the viscosity criterion for admissibility [5], [15], [7].

The critical points of a dynamical system are crucial to its study. For (2.10), a critical point is a state U_c that satisfies the Rankine-Hugoniot condition for the given state U_- and the speed s. The behavior of solutions in the neighborhood of a critical point U_c is reflected in qualitative features of solutions of the linearization of (2.10) about U_c:

$$[-s+F'(U_c)](U-U_c)=D(U_c)\dot{U}. \tag{2.11}$$

Such solutions are determined by the eigenvalues μ and corresponding eigenvectors $\hat{U}_\mu$ that satisfy

$$[-s+F'(U_c)]\hat{U}_\mu=\mu D(U_c)\hat{U}_\mu. \tag{2.12}$$

For example, $U=U_c+\Sigma_\mu c_\mu \exp(\mu\xi)\hat{U}_\mu$ when the eigenvalues are distinct. Thus the character of the critical point is determined by the eigenvalues μ.

Let us restrict ourselves now to systems of two conservation laws, so that (2.10) is a planar dynamical system. A critical point is classified as an anti-saddlepoint (i.e., a node, focus, or center) or as a saddlepoint. (This assumes that it is simple, i.e., neither eigenvalue vanishes.) To illustrate the relationship between the nature of critical points and the classification of discontinuities, we first discuss the case when $D(U_c)$ is the identity matrix; then the eigenvalues at a critical point U_c are $\mu_i=\lambda_i(U_c)-s$, $i=1,2$. This choice arises commonly in studies of viscous profiles for shock waves, but it is a degenerate case for crossing discontinuities, as we show in § 3.3.

A Lax shock wave of the first family has $s<\lambda_1(U_-)<\lambda_2(U_-)$ and $\lambda_1(U_+)<s<\lambda_2(U_+)$, so that the critical points U_- and U_+ of (2.10) are, respectively, a repelling node and a saddlepoint. Similarly, U_- and U_+ are, respectively, a saddlepoint and an attracting node in the case of a Lax shock wave of the second family. Therefore an admissible discontinuity of Lax type corresponds to a saddle-node connection. For crossing discontinuities, which are defined by (2.7) and (2.8) with $i=1$, the critical points U_- and U_+ are saddlepoints. Thus transitional shock waves correspond to saddle-saddle connections. Finally, a connecting orbit that joins a repelling node to an attracting node corresponds to a *totally compressive* shock wave:

$$\lambda_1(U_+)<s<\lambda_1(U_-), \tag{2.13}$$

$$\lambda_2(U_+)<s<\lambda_2(U_-), \tag{2.14}$$

so that the characteristics of both families impinge on the discontinuity.

In the general case where $D(U_c)$ is not a multiple of the identity matrix, the signs of $\lambda_i(U_c)-s$ do not always determine the character of a critical point U_c. Indeed, a critical point that is a node when $D(U_c)=I$ can become a focus when $D(U_c)$ is changed. However, saddlepoints are preserved provided that the determinant of $D(U_c)$ is positive; this can be demonstrated as follows. Because μ_+ and μ_- are the eigenvalues of $D(U_c)^{-1}[-s+F'(U_c)]$, their product $\mu_+\mu_-$ has the same sign as that of $(\lambda_+-s)\times(\lambda_--s)$, which is negative. Therefore μ_+ and μ_- must be real and have opposite sign. Nodes, too, are preserved if the Jacobian matrix $F'(U_c)$ is symmetric and $D(U_c)$ is symmetric and positive definite, since then $D(U_c)^{-1}[-s+F'(U_c)]$ is similar to the symmetric matrix $D(U_c)^{-1/2}[-s+F'(U_c)]D(U_c)^{-1/2}$, but this is a rather restrictive situation.

When U_- and U_+ are sufficiently close, the saddle-antisaddle nature of Lax shock waves guarantees the existence and uniqueness of a connecting orbit [6], [4] (assuming strict hyperbolicity and genuine nonlinearity). This connecting orbit is expected to be structurally stable, in the sense that the orbit persists under small perturbations of U_-, U_+, and s (subject to the jump condition (2.4)) and under changes of the viscosity matrix D. By contrast, a connecting orbit between two saddlepoints is structurally unstable.

To be precise, structural stability holds if the dynamical system is Morse-Smale [12]. For example, the system is not Morse-Smale if some critical point is nonhyperbolic (the real part of an eigenvalue is zero) or if there is a saddle-saddle connection; in these cases bifurcation is expected. In the context of conservation laws, the critical point U_+ is nonhyperbolic if $\lambda_i(U_+) = s$ for some i, i.e., at boundaries between different types of discontinuities. (These points are marked as dots in Fig. 3 below.) More generally, a boundary occurs if any critical point U_c, which corresponds to a discontinuity with speed s from U_- to U_c, has an eigenvalue μ with vanishing real part. In addition, bifurcation is expected if there is a saddle-saddle connection between two critical points. For instance, consider a 1-shock wave from U_- to U_+ such that an orbit connects another saddlepoint U_c to U_+; then U_+ can be a boundary between admissible 1-shock waves and inadmissible ones [34], [10]. (As discussed in § 5, this occurs in Fig. 8, with U_-, U_+, and U_c being points L, C, and A_1.)

We see that saddle-saddle connecting orbits have two related roles in solving Riemann problems. The first is to cause bifurcations between admissible and inadmissible shock waves. The second is to act as transitional shock waves that appear in Riemann solutions. The structural instability of saddle-saddle connections indicates that only special crossing discontinuities should have viscous profiles. Thus the class

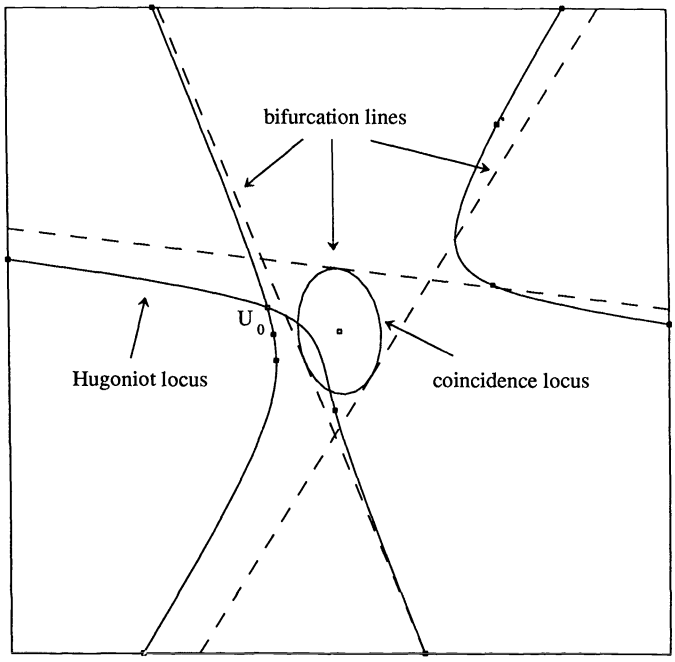


FIG. 3. *Bifurcation lines for a quadratic model, and the Hugoniot locus for a representative point* U_0. *Dots along the Hugoniot locus demarcate segments with different shock types. Also shown is the coincidence locus, where the characteristic speeds coincide; inside this curve, the system is elliptic.*

of transitional shock waves is sensitive to the precise form of the parabolic terms in (2.9): if the solution of a Riemann problem contains a transitional shock wave, then the intermediate constant states in the solution are changed if the viscosity matrix is altered. In particular, a numerical method for solving conservation laws might select waves that are not physical if it relies on artificial viscosity. (This has been emphasized to us by Schaeffer and Shearer [30].)

For completeness, we briefly describe the role of admissible totally compressive shock waves in Riemann problems. (We refer to [18] for an example of a system of conservation laws in which totally compressive waves arise; see also [32] for a discussion.) According to (2.13) and (2.14), U_- and U_+ are both nodes, so that there is an infinite number of orbits connecting them. These inequalities also imply that a totally compressive wave cannot be preceded or followed by any other wave. In other words, there is only one wave group when a totally compressive wave occurs. Thus the utility of such waves for solving Riemann problems is limited: the set of right states $U_R = U_+$ for which the Riemann solution contains a totally compressive wave is one-dimensional, comprising segments along the Hugoniot locus through $U_L = U_-$.

Let us presume that when U_R is perturbed from this set, a solution of the Riemann problem with data U_L and U_R exists and depends L^1_{loc}-continuously on U_R. Then the perturbed solution must contain a 1-wave group with a 1-shock wave on its left and a 2-wave group with a 2-shock wave on its right, with shock speeds approximately the same as that of the totally compressive wave. In the limit, as U_R moves back onto the segment of totally compressive waves, the critical points corresponding to the Lax shock waves remain joined by orbits. In particular, the dynamical system for the totally compressive wave must contain other critical points besides U_- and U_+; the repelling node U_- is also connected to a saddlepoint, and a saddlepoint connects to the attracting node U_+. From this perspective, a totally compressive wave should be regarded as containing 1- and 2-shock waves, not as a new type of shock wave.

3. Transitional shock waves in quadratic models. In this section, we examine crossing discontinuities that possess viscous profiles. These profiles correspond to saddle-saddle connecting orbits for the dynamical system (2.10). Our results are limited to systems of two conservation laws with quadratic fluxes, which are described in § 3.1. Nevertheless, we believe that the results reflect the structure of transitional shock waves for general systems of two conservation laws. (See § 5 for further discussion.)

The motivation for our analysis derives from the theory of polynomial dynamical systems in the plane: for certain quadratic systems, a saddle-saddle connecting orbit must be a straight-line segment. In § 3.2, therefore, we determine conditions under which a discontinuity (not necessarily a crossing discontinuity) possesses a viscous profile that lies along a straight line. Then, in § 3.3, we state sufficient conditions for a quadratic model to have the property that a viscous profile for a crossing discontinuity must be a straight-line segment.

3.1. Quadratic models. A quadratic model is a system of two conservation laws

$$U_t + F(U)_x = 0 \tag{3.1}$$

in which the flux is a quadratic function: writing $U = (u, v)^T$ and $F = (f, g)^T$,

$$f(u, v) = \tfrac{1}{2}(a_1 u^2 + 2b_1 uv + c_1 v^2) + d_1 u + e_1 v, \tag{3.2}$$

$$g(u, v) = \tfrac{1}{2}(a_2 u^2 + 2b_2 uv + c_2 v^2) + d_2 u + e_2 v. \tag{3.3}$$

Evidently, a quadratic flux approximates the flux for a general system of two conservation laws in the neighborhood of a point. When the linear terms are absent, the two

characteristic speeds coincide at $U = 0$; more generally, any (nondegenerate) quadratic model fails to be strictly hyperbolic somewhere in the u-v plane. Furthermore, *elliptic regions*, where the characteristic speeds are complex, may occur. The Riemann problem for quadratic models has been studied by Gomes, H. Holden, L. Holden, Isaacson, Marchesin, Paes-Leme, Plohr, Rascle, Schaeffer, Serre, Shearer, and Temple [19], [31], [35], [18], [20], [21], [13], [29], [33], [32], [10], [14], [34].

In the study of Riemann problems, the solutions of the Rankine-Hugoniot condition (2.4) play an important role. For systems of two conservation laws, it is convenient to eliminate the speed s from the Rankine-Hugoniot condition. With $U_0 = U_-$ regarded as fixed, this yields a single equation for states $U = U_+$ in the *Hugoniot locus* of U_0:

$$H_{(u_0, v_0)}(u, v) = 0, \tag{3.4}$$

where

$$H_{(u_0, v_0)}(u, v) = (u - u_0)[g(u, v) - g(u_0, v_0)] - (v - v_0)[f(u, v) - f(u_0, v_0)] \tag{3.5}$$

is the Hugoniot function. Similarly, the shock speed s is given by

$$s = \frac{(u - u_0)[f(u, v) - f(u_0, v_0)] + (v - v_0)[g(u, v) - g(u_0, v_0)]}{(u - u_0)^2 + (v - v_0)^2}. \tag{3.6}$$

An example of a Hugoniot locus is shown in Fig. 3. For quadratic models, H is a cubic polynomial in the two variables u and v. Moreover [18], the Hugoniot locus of U_0 is parameterized by angle in the polar coordinate system centered at U_0, except when the Hugoniot locus contains a line through U_0. To explain this result, we first introduce some convenient notation and terminology.

Associated to a given quadratic model are the functions α, β, γ, and $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\gamma}$, which are defined by

$$\alpha(\varphi) = \tfrac{1}{2}\{(a_2 + b_1)\cos 2\varphi + (b_2 - a_1)\sin 2\varphi + a_2 - b_1\}, \tag{3.7}$$

$$\beta(\varphi) = \tfrac{1}{2}\{(b_2 + c_1)\cos 2\varphi + (c_2 - b_1)\sin 2\varphi + b_2 - c_1\}, \tag{3.8}$$

$$\gamma(\varphi) = \tfrac{1}{2}\{(d_2 + e_1)\cos 2\varphi + (e_2 - d_1)\sin 2\varphi + d_2 - e_1\}, \tag{3.9}$$

and

$$\tilde{\alpha}(\varphi) = \tfrac{1}{2}\{(a_1 - b_2)\cos 2\varphi + (b_1 + a_2)\sin 2\varphi + a_1 + b_2\}, \tag{3.10}$$

$$\tilde{\beta}(\varphi) = \tfrac{1}{2}\{(b_1 - c_2)\cos 2\varphi + (c_1 + b_2)\sin 2\varphi + b_1 + c_2\}, \tag{3.11}$$

$$\tilde{\gamma}(\varphi) = \tfrac{1}{2}\{(d_1 - e_2)\cos 2\varphi + (e_1 + d_2)\sin 2\varphi + d_1 + e_2\}. \tag{3.12}$$

If we set $U = U_0 + R(\cos\varphi, \sin\varphi)^T$, then the Hugoniot function is

$$H_{(u_0, v_0)}(u, v) = R^2\{\tfrac{1}{2}R[\alpha(\varphi)\cos\varphi + \beta(\varphi)\sin\varphi] + \alpha(\varphi)u_0 + \beta(\varphi)v_0 + \gamma(\varphi)\}; \tag{3.13}$$

moreover,

$$s = \tfrac{1}{2}R[\tilde{\alpha}(\varphi)\cos\varphi + \tilde{\beta}(\varphi)\sin\varphi] + \tilde{\alpha}(\varphi)u_0 + \tilde{\beta}(\varphi)v_0 + \tilde{\gamma}(\varphi). \tag{3.14}$$

An angle φ is called an *asymptotic angle* when

$$\alpha(\varphi)\cos\varphi + \beta(\varphi)\sin\varphi = 0, \tag{3.15}$$

and it is called a *characteristic angle* for a given state U_0 when

$$\alpha(\varphi)u_0 + \beta(\varphi)v_0 + \gamma(\varphi) = 0. \tag{3.16}$$

The set of states U_0 satisfying (3.16) constitutes the *characteristic line* $\mathscr{L}(\varphi)$ associated with φ. At points along such a line, one of the eigenvectors has inclination angle φ. The characteristic line associated with an asymptotic angle is called a *bifurcation line* (see Fig. 3). Note that asymptotic angles and bifurcation lines depend solely on the coefficients defining the model, not on U_0.

The parameterization of Hugoniot loci is a consequence of (3.13), as shown in Proposition 3.1.

PROPOSITION 3.1 [18]. (a) *Suppose that φ is not an asymptotic angle. Then the line through U_0 at angle φ intersects the Hugoniot locus of U_0 at a state $U \neq U_0$ if and only if φ is not a characteristic angle for U_0.*

(b) *Suppose that φ is an asymptotic angle. Then the line through U_0 at angle φ intersects the Hugoniot locus of U_0 at a state $U \neq U_0$ if and only if U_0 lies on the bifurcation line associated with φ, in which case the Hugoniot locus contains this line.*

Remarks. (1) If U_0 belongs to the hyperbolic region, the Hugoniot locus through U_0 has two branches; these branches are tangent at U_0 to the right eigenvectors of $F'(U_0)$. According to (3.13), then, $(\cos\varphi, \sin\varphi)^T$ is a right eigenvector if and only if φ is a characteristic angle for U_0. By (3.14), the corresponding eigenvalue is $\lambda = \tilde{\alpha}(\varphi)u_0 + \tilde{\beta}(\varphi)v_0 + \tilde{\gamma}(\varphi)$.

(2) As demonstrated in Appendix B.1, the envelope of the characteristic lines is the coincidence locus, i.e., points where the eigenvalues coincide. In particular, bifurcation lines are tangent to the coincidence locus.

(3) The Hugoniot locus approaches infinity at the asymptotic angles (as in Fig. 3). If φ is an asymptotic angle, so is $\varphi + \pi$, and these two angles determine the same bifurcation line. Because (3.15) is a homogeneous cubic equation in $\cos\varphi$ and $\sin\varphi$, there are up to three bifurcation lines, i.e., up to six asymptotic angles $\varphi \in (-\pi, \pi]$. The viscosity angle φ is said to be *simple* if it corresponds to a simple root of the cubic.

(4) The Hugoniot locus for a state U_0 has a secondary bifurcation point if and only if U_0 lies on a bifurcation line, in which case the secondary bifurcation occurs on this line.

(5) For later purposes (see Lemma 3.4), we note that

$$\alpha(\varphi)\cos\varphi + \beta(\varphi)\sin\varphi = (-\sin\varphi, \cos\varphi)F''(0)\cdot\begin{pmatrix}\cos\varphi\\ \sin\varphi\end{pmatrix}^2, \tag{3.17}$$

$$\tilde{\alpha}(\varphi)\cos\varphi + \tilde{\beta}(\varphi)\sin\varphi = (\cos\varphi, \sin\varphi)F''(0)\cdot\begin{pmatrix}\cos\varphi\\ \sin\varphi\end{pmatrix}^2. \tag{3.18}$$

When studying viscous profiles for quadratic models, we take the viscosity matrix D to be constant; this is reasonable because quadratic models arise as expansions. Then (2.10) becomes the planar, autonomous system of ordinary differential equations

$$D\dot{U}(\xi) = -s[U(\xi) - U_-] + F(U(\xi)) - F(U_-). \tag{3.19}$$

3.2. Viscous profiles. A viscous profile for a crossing discontinuity is defined by an orbit that joins two saddlepoints. For dynamical systems that are quadratic gradients, Chicone [2] has shown that every saddle-saddle connection is a straight-line segment. With this in mind, we first construct all discontinuities (not necessarily crossing discontinuities) that have straight-line orbits. This approach yields a large class of transitional shock waves, as discussed in § 3.3.

For convenience, we use the notation $\bar{U} = \frac{1}{2}(U_+ + U_-)$ for the average of, and $\Delta U = U_+ - U_-$ for the difference between, two states U_+ and U_-. The construction relies on an obvious property of quadratic functions.

LEMMA 3.2. *Suppose Q is a quadratic function such that $Q(U_+) = Q(U_-)$. Then*

$$Q(\bar{U}+\rho\Delta U) = Q(\bar{U}) + \tfrac{1}{2}\rho^2 Q''(\bar{U})\cdot(\Delta U)^2. \tag{3.20}$$

PROPOSITION 3.3. *Let F be quadratic, and suppose that s, U_-, and $U_+ \neq U_-$ satisfy the Rankine-Hugoniot condition* (2.4). *Then the straight-line segment between U_- and U_+ is a connecting orbit for* (3.19) *if and only if there is a constant $\mu \neq 0$ such that*

$$\mu D\Delta U = \tfrac{1}{2}F''(0)\cdot(\Delta U)^2. \tag{3.21}$$

The orbit is traversed from U_- to U_+ if and only if $\mu < 0$.

Proof. An orbit connecting U_- and U_+ along a line takes the form

$$U(\xi) = \bar{U} + \rho(\xi)\Delta U \tag{3.22}$$

with $-\frac{1}{2} < \rho(\xi) < \frac{1}{2}$. If the quadratic function Q in Lemma 3.2 is defined by $Q(U) = -s(U-U_-) + F(U) - F(U_-)$, then the dynamical system (3.19) becomes $D\dot{U} = Q(U)$, i.e.,

$$\dot{\rho} D\Delta U = Q(\bar{U}) + \tfrac{1}{2}\rho^2 Q''(\bar{U})\cdot(\Delta U)^2. \tag{3.23}$$

But $0 = Q(U_-) = Q(\bar{U}) + \frac{1}{8}Q''(\bar{U})\cdot(\Delta U)^2$ and $Q''(\bar{U}) = F''(0)$, so

$$\dot{\rho} D\Delta U = \tfrac{1}{2}(\rho^2 - \tfrac{1}{4})F''(0)\cdot(\Delta U)^2. \tag{3.24}$$

This equation is satisfied if (3.21) holds and

$$\dot{\rho} = \mu(\rho^2 - \tfrac{1}{4}). \tag{3.25}$$

Provided that $\mu \neq 0$, (3.25) has a solution with ρ varying between $-\frac{1}{2}$ and $\frac{1}{2}$. Conversely, if ρ parameterizes a connecting orbit along a straight line, (3.24) shows that (3.25) must hold for some $\mu \neq 0$, and therefore that (3.21) is satisfied. The parameter ρ increases from $-\frac{1}{2}$ to $\frac{1}{2}$, i.e., the orbit is traversed from U_- to U_+, if and only if $\mu < 0$. □

The quantity μ is related to the eigenvalues for the linearized differential equations at the critical points, as we now show. Suppose that the straight-line segment from U_- to U_+ is an orbit; then it must coincide with an unstable manifold for U_- and a stable manifold for U_+. Therefore, by (2.12),

$$[-s + F'(U_\pm)]\Delta U = \mu_\pm D\Delta U \tag{3.26}$$

with $\mu_+ \leqq 0 \leqq \mu_-$. Subtracting these two equations yields

$$[F'(U_+) - F'(U_-)]\Delta U = (\mu_+ - \mu_-)D\Delta U. \tag{3.27}$$

Since $F'(U_+) - F'(U_-) = F''(0)\Delta U$, we obtain (3.21) with $\mu = \frac{1}{2}(\mu_+ - \mu_-) \leqq 0$. Moreover, adding (3.26) shows that

$$[-s + F'(\bar{U})]\Delta U = \tfrac{1}{2}(\mu_+ + \mu_-)D\Delta U. \tag{3.28}$$

According to the midpoint rule for quadratic models [18], the left-hand side vanishes when U_-, U_+, and s satisfy the Rankine-Hugoniot condition

$$s\Delta U = F(U_+) - F(U_-) = \int_{-1/2}^{1/2} F'(\bar{U} + \rho\Delta U)\Delta U\, d\rho = F'(\bar{U})\Delta U. \tag{3.29}$$

Consequently, $\mu_+ + \mu_- = 0$, so that $\mu = \mu_+ = -\mu_-$.

As the next lemma shows, solutions of (3.21) are related to the asymptotic angles for the quadratic model with the flux function $D^{-1}F$. Because these angles are determined by the viscosity matrix D as well as by F, we call them *viscosity angles.* Also, the characteristic line $\mathscr{L}(\varphi)$ associated with a viscosity angle is called a *viscosity line.*

LEMMA 3.4. *Let U_+ lie on the line through U_- at angle φ, with $U_+ \neq U_-$. Then* (3.21) *holds for some μ if and only if φ is a viscosity angle.*

Proof. Let $\Delta U = R(\cos\varphi, \sin\varphi)^T$. Then (3.21) holds for some μ if and only if

$$0 = (-\sin\varphi, \cos\varphi) D^{-1} F''(0) \cdot \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix}^2. \tag{3.30}$$

Comparing this with (3.17), we see that (3.21) holds if and only if φ is an asymptotic angle for the quadratic model with flux $D^{-1}F$. □

The existence of straight-line orbits depends also on the eigenvalue μ. To determine a formula for μ, suppose that φ is a viscosity angle and that U_+ lies on the line through U_- at angle φ, say $U_+ = U_- + R(\cos\varphi, \sin\varphi)^T$. Then (3.21) and (3.18) imply that

$$\mu = \tfrac{1}{2} R[\tilde{\alpha}_D(\varphi)\cos\varphi + \tilde{\beta}_D(\varphi)\sin\varphi]; \tag{3.31}$$

here $\tilde{\alpha}_D$ and $\tilde{\beta}_D$ are the functions associated with $D^{-1}F$ that are defined (for F) in (3.10) and (3.11). A viscosity angle φ is said to be *exceptional* if $\tilde{\alpha}_D(\varphi)\cos\varphi + \tilde{\beta}_D(\varphi)\sin\varphi = 0$: no straight-line profiles are possible at an exceptional viscosity angle because $\mu = 0$. Otherwise, if φ is not exceptional, then the sign of μ is determined by the sign of R. Thus $\mu < 0$ on a particular open ray with respect to U_-. For simplicity, we say that U_+ is *correctly oriented* with respect to U_- along the viscosity line if it lies in this ray.

The construction of discontinuities with straight-line profiles proceeds as follows. The line at each viscosity angle is drawn through U_0, and its intersection U with the Hugoniot locus through U_0 is found; then U and U_0 are joined by a profile along this direction. Of course, this intersection might not exist; the precise conditions are consequences of the characterization of Hugoniot loci given in Proposition 3.1.

THEOREM 3.5. *Assume that the viscosity matrix D is invertible, and consider a fixed viscosity angle φ for D that is not exceptional. Let $\mathcal{L}(\varphi)$ be the viscosity line associated with φ.*

(a) *Suppose that φ is not an asymptotic angle. Then U_0 is connected to some $U \neq U_0$ on the Hugoniot locus of U_0 by a connecting orbit lying along a straight line at angle φ if and only if $U_0 \notin \mathcal{L}(\varphi)$. In this case, the corresponding state U is unique.*

(b) *Suppose that φ is an asymptotic angle. (Thus $\mathcal{L}(\varphi)$ is a bifurcation line.) Then U_0 is connected to some $U \neq U_0$ on the Hugoniot locus of U_0 by a connecting orbit lying along a straight line at angle φ if and only if $U_0 \in \mathcal{L}(\varphi)$. In this case, the corresponding states U comprise all of $\mathcal{L}(\varphi)$.*

The connecting orbit is traversed from U_0 to U if and only if U is correctly oriented with respect to U_0.

Part (a) of this theorem is illustrated in Fig. 4, which shows three discontinuities that possess straight-line profiles. Part (b) is nongeneric, but it arises when D is a multiple of the identity, which is the simplest choice. Bifurcations of saddle-saddle connections under cubic perturbations has been studied in the case $D = I$ in [36].

3.3. Saddle-saddle connecting orbits. Theorem 3.5 characterizes when a discontinuity in a quadratic model possesses a straight-line profile. To apply it to solving Riemann problems, we must account for the wave type of the discontinuity, i.e., the relationships between the propagation speed and the characteristic speeds on the two sides of the discontinuity. Indeed, Theorem 3.5 allows a discontinuity with a straight-line profile to be of either Lax or crossing type. Lax discontinuities with straight-line profiles, however, constitute only a small subset of the Lax discontinuities with viscous profiles. This is because saddle-node connecting orbits are structurally stable. By

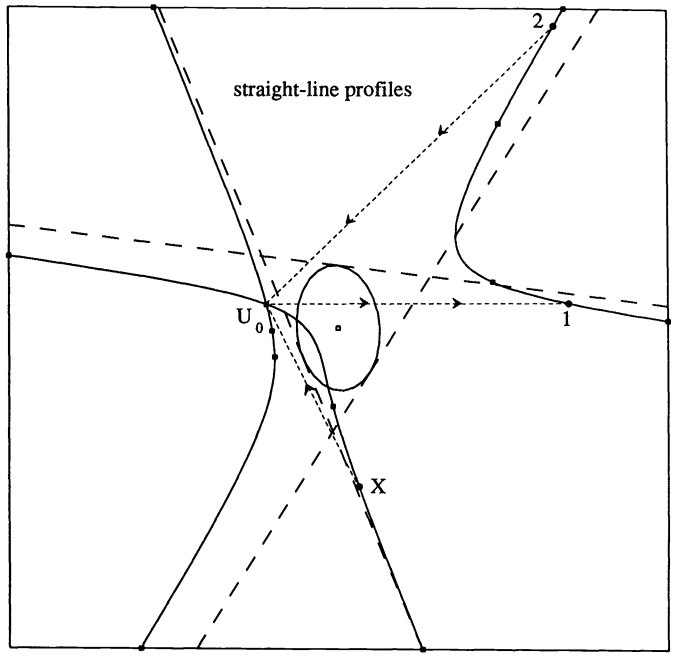


FIG. 4. *Discontinuities with straight-line profiles. The dashed lines are drawn through U_0 at the viscosity angles; arrows indicate the direction of the connecting orbit. The discontinuity corresponding to point X is of crossing type, while point 1 is the right state of a 1-shock and point 2 is the left state of a 2-shock.*

contrast, profiles for crossing discontinuities, which correspond to saddle-saddle connecting orbits, do not persist when the discontinuity suffers a generic perturbation.

In this section, we describe the set of crossing discontinuities with straight-line profiles. Furthermore, for certain classes of quadratic models, we show that the only crossing discontinuities with viscous profiles take the form constructed in Theorem 3.5. Within these classes, therefore, the set of points (U_L, s, U_R) corresponding to transitional shock waves has codimension 3, whereas the set corresponding to Lax shock waves has codimension 2.

Assume that the hypotheses of Theorem 3.5 hold. In both cases (a) and (b), any point U_0 in a certain set is connected to some $U \neq U_0$ by a straight-line profile at angle φ. We define the *transitional region* for the viscosity angle φ to be the subset of points U_0 for which at least one of these discontinuities is of crossing type. Thus the transitional region is defined by

$$\lambda_1(U_0) < s < \lambda_2(U_0), \tag{3.32}$$

$$\lambda_1(U) < s < \lambda_2(U), \tag{3.33}$$

and the requirement of being correctly oriented. In the situation of part (a), the transitional region is an open subset of the plane, and to each of its points corresponds a unique admissible crossing discontinuity for φ. In fact, the transitional region is a wedge, as we show presently. This generic case is illustrated in Fig. 5. Similarly, for part (b), the transitional region is a ray of the bifurcation line $\mathscr{L}(\varphi)$, and to each point in this set corresponds an open interval of admissible crossing discontinuities.

The precise form of the boundary of the transitional region is determined as follows. We consider a particular viscosity angle φ and allow the point U_0 to vary;

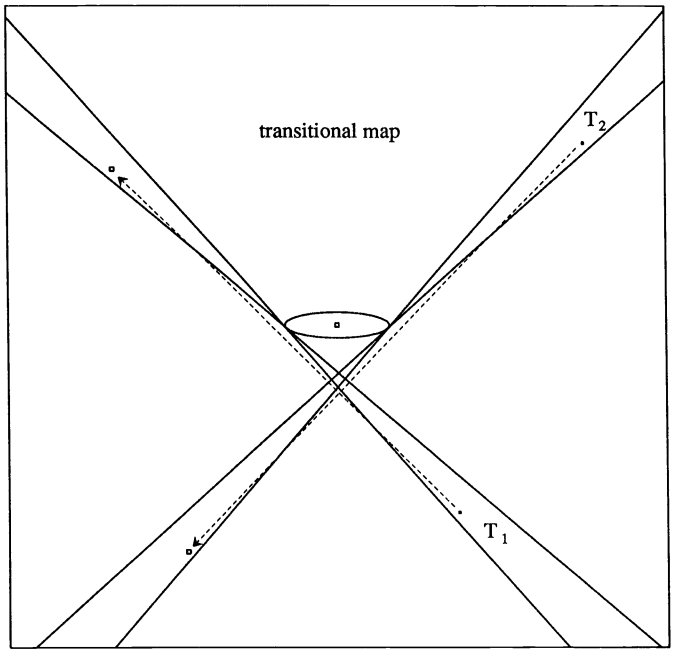


FIG. 5. *Transitional regions. A point in region T_1 or T_2 is the left state for a transitional shock wave. The corresponding right state is given by the transitional map indicated by the dashed lines, which lie at viscosity angles. Two of the six viscosity angles do not lead to transitional shock waves, so that there is no corresponding transitional region.*

the corresponding point $U = U_0 + R(\cos\varphi, \sin\varphi)^T$ is joined to U_0 by a straight-line profile. The boundary of the transitional region consists of points U_0 for which some of the inequalities in (3.32) and (3.33) become equalities, so that the speed of the shock coincides with the characteristic speed at either the left or the right state: $\det[-s + F'(U_c)] = 0$, with U_c being U_0 or U. By (3.13) and (3.14) each of these two equations is quadratic in u_0 and v_0. Points U_0 on the viscosity line satisfy these equations, so that each factors into a product of linear polynomials, i.e., each solution set is a pair of crossed lines. Note, however, that the solutions on the viscosity line are irrelevant because $R = 0$ along this line. We conclude that the boundary of the transitional region is contained in two crossed lines. One ray of each line corresponds to profiles that are correctly oriented, so that the transitional region is a wedge. In the case of part (b), the wedge collapses to a single ray.

We now turn our attention to the question of whether all saddle-saddle connections are straight-line segments, so that the transitional shock waves constructed above are the only ones. This is true for certain classes of quadratic models and certain choices for the viscosity matrix. To prove this, we draw upon results for quadratic planar dynamical systems, of which (3.19) is an example. Such systems have been studied in connection with Hilbert's 16th problem (see, e.g., Chicone and Jinghuang [3]).

Some of the results we describe concern quadratic models that are strictly hyperbolic except at $U = 0$. Schaeffer and Shearer [31] have classified such models; they have shown that a linear change of dependent variables U brings the system of conservation laws to a normal form in which the flux is a gradient, i.e., $F(U) = C'(U)$. The qualitative structure of solutions falls into four categories, corresponding to four regions in parameter space, which are labeled Cases I–IV.

When the flux is a gradient, the vector field on the right-hand side of the dynamical system (3.19) is also a gradient: viscous profiles satisfy

$$D\dot{U}(\xi) = G'(U(\xi)), \tag{3.34}$$

where G is defined by

$$\begin{aligned} G(u, v) = C(u, v) - C(u_-, v_-) - \tfrac{1}{2}s[(u-u_-)^2 + (v-v_-)^2] \\ - f(u_-, v_-)(u-u_-) - g(u_-, v_-)(v-v_-). \end{aligned} \tag{3.35}$$

In particular, when the viscosity matrix D is a multiple of the identity matrix, the dynamical system (3.34) is a quadratic gradient system in the plane. The following result of Chicone [2] bears directly on our application.

THEOREM 3.6 (Chicone [2]). *For a quadratic gradient system in the plane, an orbit connecting two saddlepoints is a straight-line segment.*

Note that where D is a multiple of the identity matrix, viscosity angles coincide with asymptotic angles. Therefore part (b) of Theorem 3.5 yields the following.

COROLLARY 3.7. *Suppose that the flux for a quadratic model is a gradient. Then if the viscosity matrix D is a multiple of the identity matrix, any viscous profile for a crossing discontinuity must lie along a straight line. Furthermore, a crossing discontinuity connecting U_- to U_+ has a viscous profile if and only if U_- and U_+ both lie on the same bifurcation line and U_+ is correctly oriented with respect to U_-.*

For general choices of D, however, viscosity angles differ from asymptotic angles, so that part (a) of Theorem 3.5 applies instead. The class of discontinuities with straight-line profiles then takes a different form: for each state U_- in a certain open set, there is a finite set of states U_+ corresponding to such discontinuities. In this sense, the case where D is a multiple of the identity matrix is not representative of the generic case. Extension of Corollary 3.7 to more general viscosity matrices requires generalizing Chicone's theorem. Recently, Gomes [9], [11] has proved one such generalization.

THEOREM 3.8 (Gomes [9], [11]). *Consider a quadratic dynamical system in the plane with more than two critical points at infinity; assume that the total topological index of these critical points is neither 4 nor 6. Then any orbit connecting two saddlepoints is a straight-line segment.*

As shown in Appendix A, critical points at infinity for the quadratic system (3.19) occur precisely at the viscosity angles. Furthermore, if the viscosity angle φ is simple and nonexceptional, the topological index $\mathrm{Ind}\,(\varphi)$ of the corresponding critical point at infinity is

$$\mathrm{Ind}\,(\varphi) = -\mathrm{sgn}\left[\frac{d}{d\varphi}\{\alpha_D(\varphi)\cos\varphi + \beta_D(\varphi)\sin\varphi\}\cdot\{\tilde{\alpha}_D(\varphi)\cos\varphi + \tilde{\beta}_D(\varphi)\sin\varphi\}\right].$$

On the basis of these results and part (a) of Theorem 3.5, we can extend Corollary 3.7 as follows.

COROLLARY 3.9. *Consider a quadratic model together with a viscosity matrix D. Suppose that there are six nonexceptional viscosity angles, and that $\sum_\varphi \mathrm{Ind}\,(\varphi) \neq 6$. Then any viscous profile for a crossing discontinuity must lie along a straight line.*

Note that $\sum_\varphi \mathrm{Ind}\,(\varphi)$ is determined by D and the homogeneous quadratic part F_2 of F. Suppose that the quadratic model with flux $D^{-1}F_2$ is strictly hyperbolic except at $U = 0$, so that it falls into the Schaeffer-Shearer classification. Then there are six viscosity angles in Cases I–III and two in Case IV. Moreover, $\sum_\varphi \mathrm{Ind}\,(\varphi)$ is 6 in Case I and 2 in Cases II–IV. Therefore the hypotheses of Corollary 3.9 are satisfied in Cases II and III.

We emphasize, however, that for some quadratic models, saddle-saddle connecting orbits need not be straight-line segments. Azevedo [1] has given such an example: for a certain quadratic model with an elliptic region, U_- and U_+ can be chosen so that there is a curved saddle-saddle connecting orbit from U_+ to U_- even though the straight line from U_- to U_+ is not invariant. In this example, $D = I$ and the homogeneous quadratic part of F falls into Case I.

4. Transitional rarefaction waves. In this section, we study transitional rarefaction waves for systems of two conservation laws. Such a wave arises when an integral curve of family 2 is followed by an integral curve of family 1 (in the direction of increasing characteristic speed). Of necessity, the two characteristic speeds coincide at the point where these curves join. Suppose that the set of states U at which $\lambda_1(U) = \lambda_2(U)$ forms a smooth curve separating a hyperbolic region from an elliptic region. (Models for which the coincidence locus separates two regions of strict hyperbolicity are possible also; one example is discussed in Appendix B.2.) Figure 6 illustrates four conceivable configurations of rarefaction curves in the vicinity of this curve. Transitional rarefaction waves occur in Fig. 6b, d, while in Fig. 6a, c they do not. Note that solutions of Riemann problems are not unique if the configuration resembles Fig. 6b: both U_M and U_T serve as middle states in solutions of the Riemann problem with data U_L and U_R.

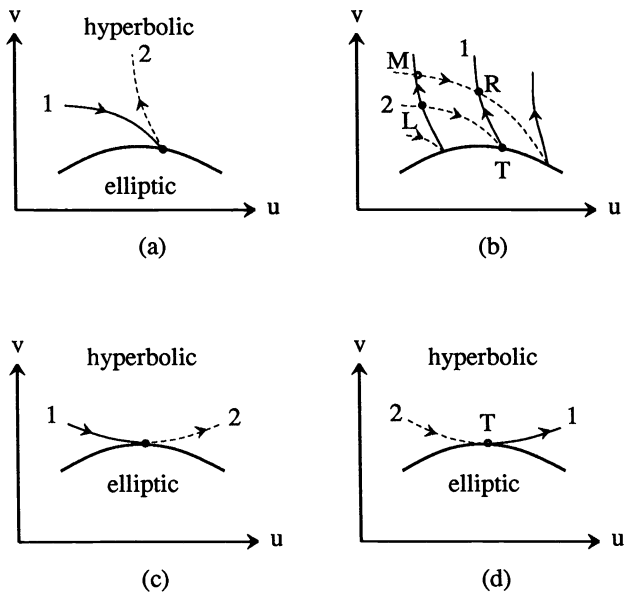


FIG. 6. *Rarefaction curves near an elliptic region. In* (a) *and* (c) *there are no transitional rarefaction waves; in* (b) *and* (d), *the curves through points such as T correspond to transitional rarefaction waves. For generic fluxes, points such as T are isolated, so that configuration* (b), *which causes nonuniqueness, does not arise.*

In the following, we present a detailed picture of the behavior of integral curves near the boundary of an elliptic region. We employ the approach of Palmeira [28], who studied integral curves for quadratic models with compact elliptic regions. One consequence of this analysis is that the configuration of Fig. 6a is generic, whereas points such as T in Fig. 6b, d are isolated points for generic flux functions. Thus the situation of Fig. 6b, in which solutions of Riemann problems are not unique, does not occur generically. We emphasize that the present results are not restricted to quadratic models: the flux functions need satisfy only smoothness and genericity assumptions.

A rarefaction wave of a given family i is constructed using a curve in state space such that its tangent $\dot{U}$ is a right eigenvector:

$$F'(U)\dot{U} = \lambda_i(U)\dot{U}. \tag{4.1}$$

In the neighborhood of a point of strict hyperbolicity, where the eigenvalues are real and distinct, such a curve is constructed by choosing a smooth field r_i of right eigenvectors and integrating the differential equation

$$\dot{U} = r_i(U). \tag{4.2}$$

This choice is not generally possible, however, near a point where eigenvalues coincide.

To address this problem, we adopt a global geometric approach. Note that the matrix $F'(U) - \lambda_i(U)$ has rank $n-1$ in the strictly hyperbolic region, so that (4.1) constrains $\dot{U}$ to lie in a line. From this perspective, we can construct rarefaction waves using integral curves of line fields. There are n distinct line fields defined throughout the strictly hyperbolic region. As we explain below, however, two of these line fields join smoothly at the boundary of this region, where eigenvalues coincide. In fact, when $n = 2$, the line fields may be regarded as projections of a single line field that is defined on a larger space.

Remark. A line field in an n-dimensional manifold can be specified by intersecting $n-1$ fields of tangent hyperplanes, so long as they are linearly independent. It proves convenient to regard any $n-1$ fields of tangent hyperplanes as defining a line field; in this case, points where the hyperplanes are not independent are called *critical points* of the line field.

Let us now restrict ourselves to the case of two conservation laws. Recall that lines through the origin in $\mathbf{R}^2$ form the one-dimensional real projective space RP^1. A point in RP^1 may be identified with a normalized vector $\pm(\cos\varphi, \sin\varphi)^T$, modulo sign; as a coordinate for RP^1, therefore, we take $\varphi \in (-\pi/2, \pi/2]$ to correspond to the line through the origin that lies at angle φ. In these terms, a line field on $\mathbf{R}^2$ associates a point in RP^1 to each point in $\mathbf{R}^2$.

Following Palmeira, we introduce the space $\mathcal{P} = \mathbf{R}^2 \times RP^1$ of lines through points $U = (u, v)^T \in \mathbf{R}^2$. The map $(U, \varphi) \mapsto U$ projects $\mathcal{P}$ onto $\mathbf{R}^2$, making $\mathcal{P}$ into a fiber bundle. A line field on $\mathbf{R}^2$ may be regarded as associating a point $(U, \varphi(U)) \in \mathcal{P}$ to each point $U \in \mathbf{R}^2$. An integral curve $\xi \mapsto U(\xi)$ of such a line field is the projection of the curve defined by $\xi \mapsto (U(\xi), \varphi(U(\xi)))$; this curve in $\mathcal{P}$ is called the lift of the curve in $\mathbf{R}^2$. By definition, $\dot{U}$ lies at angle φ, so that $-\sin\varphi\,\dot{u} + \cos\varphi\,\dot{v} = 0$. This means that the vector $(\dot{U}, \dot{\varphi})$ is constrained to lie in the plane defined by the differential expression

$$-\sin\varphi\, du + \cos\varphi\, dv = 0. \tag{4.3}$$

In other words, the tangent vector of the lifted curve lies in the tangent plane at (U, φ) given by (4.3).

The line fields of interest to us are associated with eigenvectors of the Jacobian derivative matrix for the system of conservation laws. Let (U, φ) be a point in $\mathcal{P}$; then $(\cos\varphi, \sin\varphi)^T$ is a right eigenvector of $F'(U)$ if and only if

$$\mathcal{F}(U, \varphi) = (-\sin\varphi, \cos\varphi)F'(U)\begin{pmatrix}\cos\varphi \\ \sin\varphi\end{pmatrix} \tag{4.4}$$

is zero. Thus we are led to study the surface $\mathcal{F} = 0$ in $\mathcal{P}$; we call it the *characteristic surface* for the system of conservation laws. The plane tangent to $\mathcal{P}$ defined by (4.3), when intersected with the plane tangent to the characteristic surface, defines a line field on this surface. The integral curves of this line field project onto solutions of (4.1) by virtue of (4.3). A portion of a characteristic surface is depicted in Fig. 7.

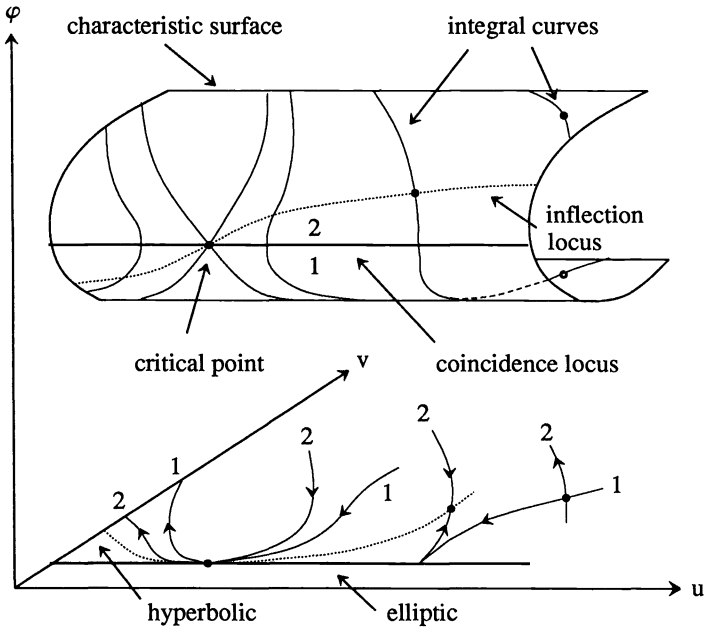


FIG. 7. *A portion of the characteristic surface in the space $\mathcal{P}$. The surface folds along the coincidence locus, which projects onto the boundary of the elliptic region. Typical integral curves are drawn on the surface; their projections onto the u-v plane are rarefaction curves.*

Recall that (U, φ) is a *regular point* of $\mathcal{F}$ if $(\mathcal{F}_u, \mathcal{F}_v, \mathcal{F}_\varphi) \neq (0, 0, 0)$ at (U, φ), and that v is a *regular value* of $\mathcal{F}$ if all points (U, φ) for which $\mathcal{F}(U, \varphi) = v$ are regular. The characteristic surface is a smooth, two-dimensional manifold in a neighborhood of a regular point, so that the whole characteristic surface is a smooth manifold provided that $v = 0$ is a regular value. Sard's theorem implies that regular values are generic if the flux F is smooth, but in general the characteristic surface might have singularities and self-intersections. The global structure of the characteristic surface is described for some examples in Appendix B.

In working with general conservation laws, it is convenient to represent the 2×2 matrix $F'(U)$ as

$$F'(U) = \begin{pmatrix} d+a & b+c \\ b-c & d-a \end{pmatrix}, \tag{4.5}$$

in terms of the functions a, b, c, and d of U (cf. [31]). Thus

$$\mathcal{F} = b \cos 2\varphi - a \sin 2\varphi - c. \tag{4.6}$$

Furthermore, in analogy with (4.4), we define the function λ on $\mathcal{P}$ by

$$\begin{aligned} \lambda(U, \varphi) &= (\cos \varphi, \sin \varphi) F'(U) \begin{pmatrix} \cos \varphi \\ \sin \varphi \end{pmatrix} \\ &= a \cos 2\varphi + b \sin 2\varphi + d. \end{aligned} \tag{4.7}$$

As seen from (4.4) and (4.7), if a point (U, φ) lies on the characteristic surface $\mathcal{F} = 0$, then $\lambda(U, \varphi)$ is the eigenvalue of $F'(U)$ for the right eigenvector $(\cos \varphi, \sin \varphi)^T$. Moreover, the following relations are easily verified:

$$(\lambda - d)^2 + (\mathcal{F} + c)^2 = a^2 + b^2, \tag{4.8}$$

$$-\tfrac{1}{2}\mathscr{F}_\varphi = \lambda - d, \tag{4.9}$$

$$-\tfrac{1}{4}\mathscr{F}_{\varphi\varphi} = \mathscr{F} + c = \tfrac{1}{2}\lambda_\varphi. \tag{4.10}$$

We define the *coincidence locus* to comprise points (U, φ) on the characteristic surface at which $\lambda_1(U) = \lambda_2(U)$ (see Fig. 7). The next result characterizes this locus.

PROPOSITION 4.1. *The coincidence locus comprises points satisfying* $\mathscr{F} = 0$ *and* $\mathscr{F}_\varphi = 0$. *The following are equivalent at a coincidence point, provided that it is a regular point of* $\mathscr{F}$: (i) *the projection of the characteristic surface is a fold*; (ii) $\lambda_\varphi \neq 0$; (iii) $c \neq 0$; *and* (iv) $(a, b) \neq (0, 0)$.

Proof. The characteristic speeds coincide if and only if the discriminant discrm $F'(U) = 4(a^2 + b^2 - c^2)$ vanishes. According to (4.8), $a^2 + b^2 - c^2 = (\lambda - d)^2$ on the surface $\mathscr{F} = 0$, so that by (4.9), coincidence occurs precisely when $\mathscr{F}_\varphi = 0$.

In particular, the projection $(U, \varphi) \mapsto U$, restricted to the characteristic surface, is singular at a coincidence point. This singularity is of fold type when $\mathscr{F}_{\varphi\varphi} \neq 0$. By (4.10) with $\mathscr{F} = 0$, conditions (i)-(iii) are equivalent, and because $a^2 + b^2 = c^2$ at a coincidence point, conditions (iii) and (iv) are equivalent. □

Accordingly, the coincidence locus is a smooth curve through those coincidence points for which the matrix

$$\begin{pmatrix} \mathscr{F}_u & \mathscr{F}_v & 0 \\ \mathscr{F}_{\varphi u} & \mathscr{F}_{\varphi v} & \mathscr{F}_{\varphi\varphi} \end{pmatrix} \tag{4.11}$$

has rank 2. A sufficient condition is that the coincidence point be a regular foldpoint. Vectors tangent to the coincidence locus belong to the kernel of this matrix. Note that a tangent vector at a regular foldpoint cannot be vertical (i.e., have vanishing u and v components).

Corresponding to the integral curves in $\mathbf{R}^2$ used to construct rarefaction waves are lifted curves lying in $\mathscr{P}$. In addition to satisfying (4.3), these lifted curves belong to the characteristic surface $\mathscr{F} = 0$, so that $d\mathscr{F} = 0$ along them. Therefore we consider integral curves of the line field in $\mathscr{P}$ given by

$$\mathscr{F}_u\, du + \mathscr{F}_v\, dv + \mathscr{F}_\varphi\, d\varphi = 0, \tag{4.12}$$

$$-\sin\varphi\, du + \cos\varphi\, dv = 0. \tag{4.13}$$

If such an integral curve starts at a point in the characteristic surface, then it lies entirely within the characteristic surface, as shown in Fig. 7.

Integral curves of (4.12) and (4.13) can be obtained locally by integrating the differential equations

$$\dot{u} = -\mathscr{F}_\varphi \cos\varphi, \tag{4.14}$$

$$\dot{v} = -\mathscr{F}_\varphi \sin\varphi, \tag{4.15}$$

$$\dot{\varphi} = \mathscr{F}_u \cos\varphi + \mathscr{F}_v \sin\varphi. \tag{4.16}$$

Indeed, this local vector field satisfies (4.12) and (4.13), and it vanishes when these two equations are linearly dependent. (Note, however, that (4.14)-(4.16) do not define a global vector field on $\mathscr{P}$: they are not invariant under the map $\varphi \mapsto \varphi + \pi$.) Thus critical points of the line field occur precisely when $\mathscr{F}_\varphi = 0$ and

$$\mathscr{F}_u \cos\varphi + \mathscr{F}_v \sin\varphi = 0. \tag{4.17}$$

We say that a point on the coincidence locus is *critical* whenever (4.17) holds. Such points play a significant role in determining the structure of integral curves on the characteristic surface [28]. We emphasize, however, that for generic choices of the flux functions in the conservation laws, critical points will be isolated points on the

coincidence locus (cf. Appendix B.2). At generic points on the coincidence locus, $\dot{u} = 0$, $\dot{v} = 0$, and $\dot{\varphi} \neq 0$; therefore the integral curve is vertical, and the projected integral curve has a cusp (see Fig. 7).

Also of importance is the *family* to which a point on the characteristic surface belongs. Noting that coincidence of eigenvalues occurs when $\lambda = d$, we define the 1-family region $\mathcal{P}_1$ in $\mathcal{P}$ to comprise points for which $\lambda < d$; similarly, $\lambda > d$ in the 2-family region $\mathcal{P}_2$. This definition is appropriate because of the following consideration. Suppose that (U, φ_1) and (U, φ_2) are two points in the characteristic surface that project onto the same point U; suppose further that $(U, \varphi_1) \in \mathcal{P}_1$ and $(U, \varphi_2) \in \mathcal{P}_2$. Then $\lambda(U, \varphi_1) < d(U) < \lambda(U, \varphi_2)$. Consequently, $\lambda(U, \varphi_1) = \lambda_1(U)$ and $\lambda(U, \varphi_2) = \lambda_2(U)$, while $(\cos \varphi_1, \sin \varphi_1)^T$ and $(\cos \varphi_2, \sin \varphi_2)^T$ are the corresponding right eigenvectors. If the system of conservation laws is strictly hyperbolic, then the characteristic surface consists of two distinct sheets, belonging to the regions $\mathcal{P}_1$ and $\mathcal{P}_2$. More generally, these sheets join along the coincidence locus, as in Fig. 7.

With this notion of family, we can state the main result of this section.

THEOREM 4.2. *Consider a regular foldpoint on the coincidence locus, and suppose that it is not critical. Then the integral curve through this point crosses the coincidence locus transversally, and its projection, a rarefaction curve, has a cusp. Moreover, if the integral curve is followed in the direction of increasing eigenvalue, then it leads from the 1-family region to the 2-family region.*

Proof. According to (4.14)-(4.16), the integral curve is vertical at such a point: $\dot{u} = 0$, $\dot{v} = 0$, and $\dot{\varphi} \neq 0$. Therefore the integral curve is transverse to the coincidence locus, whose tangent is not vertical. Also, the integral curve leads from $\mathcal{P}_1$ to $\mathcal{P}_2$ because the derivative of $\lambda - d$ along the curve is $\lambda_\varphi \dot{\varphi} \neq 0$. □

Transitional rarefaction waves arise only when an integral curve leads from the 2-family region into the 1-family region when traversed with increasing eigenvalue. By Theorem 4.2 this is possible only for integral curves through special points on the coincidence locus: critical points, where (4.17) holds; nonfoldpoints, where $\mathcal{F}_{\varphi\varphi} = 0$ (which usually correspond to cusps in the coincidence locus, as projected onto state space); and irregular points, where $(\mathcal{F}_u, \mathcal{F}_v) = (0, 0)$ (at which the characteristic surface need not be a manifold). Such singular points, being characterized by extra functional conditions, generically are isolated points on the coincidence locus.

COROLLARY 4.3. *For generic choices of the flux functions that define the system of conservation laws, transitional rarefaction waves arise only from integral curves through isolated points on the coincidence locus.*

Rarefaction waves correspond to segments of integral curves along which the characteristic speed does not decrease. To reflect this, the integral curves shown in Fig. 7 have been oriented according to the variation of λ. The orientation changes at points where $\lambda_i'(U) r_i(U) = 0$, which are called inflection points by analogy with scalar conservation laws. In the present geometric framework, the inflection locus is defined to be points on the characteristic surface for which $d\lambda = 0$ in the direction of the integral curve, i.e.,

$$\det \begin{pmatrix} \mathcal{F}_u & \mathcal{F}_v & \mathcal{F}_\varphi \\ -\sin \varphi & \cos \varphi & 0 \\ \lambda_u & \lambda_v & \lambda_\varphi \end{pmatrix} = 0. \tag{4.18}$$

The following result helps elucidate the behavior of integral curves near critical points.

PROPOSITION 4.4. *A regular foldpoint on the coincidence locus is a point of inflection if and only if it is a critical point.*

Proof. When $\mathcal{F}_\varphi = 0$, (4.18) becomes $\lambda_\varphi [\mathcal{F}_u \cos \varphi + \mathcal{F}_v \sin \varphi] = 0$. □

As illustrated in Fig. 7, the inflection locus passing through the critical point permits a transitional rarefaction to occur even though neighboring integral curves lead from family 1 to family 2.

The constructions of this section have been applied by Palmeira to quadratic conservation laws [28]. (See Appendix B.1 for another presentation of these computations.) Furthermore, Palmeira shows that the results obtained for quadratic models are stable under perturbations of the flux functions (in the C^3 Whitney topology). This is a first step in proving stability of solutions of Riemann problems with respect to changes in the conservation laws.

5. The role of transitional waves in solving Riemann problems. A solution of a Riemann problem consists of a sequence of rarefaction waves and discontinuities. For various systems (e.g., [35], [13], [9], [10], [14], [34]), the general Riemann problem cannot be solved globally if only Lax shock waves with viscous profiles are used: for certain left states, there are regions of right states for which there is no solution. In this section, we explain how transitional waves can be used to overcome this difficulty. The transitional shock waves define a certain map in state space; for simplicity, we describe this map for systems of two conservation laws, although a straightforward generalization can be made to systems of arbitrary size. We begin by recalling the classical method for constructing local solutions of Riemann problems [25], [26].

The 1-*wave curve* based upon a state U_0 consists of those states to which U_0 can be joined by a succession of 1-waves; a similar definition holds for 2-waves. Near U_0, each wave curve consists of a shock branch joined to a rarefaction branch (if U_0 is not on the inflection locus); an eigenvector of $F'(U_0)$ is tangent to these branches at U_0. To solve the Riemann problem near a left state U_L, the standard construction is to build the 1-wave curve based on U_L, and then to build the 2-wave curve based on each middle state U_M on the 1-wave curve. This is illustrated in Fig. 8 in the vicinity of U_L. (In Figs. 8 and 9, 1-wave curves are thicker than 2-wave curves, with solid curves being rarefaction curves and dashed curves being shock curves.) In this way, all neighboring states U_R are joined to U_L by a 1-wave followed by a 2-wave. To carry out this procedure globally, detached branches of the wave curves must also be used. For example, a state U_R in the upper left corner of Fig. 8 is reached by a nonlocal 1-shock wave, U_M being a point on the Hugoniot branch above point C, followed by a 2-wave. As mentioned above, however, there might be right states that are not reached in this manner, such as points in the strip bounded by the 2-wave curves through B and C. To complete the solution, transitional waves are employed.

First we describe the use of transitional shock waves. Based on part (a) of Theorem 3.5 for quadratic models, the class of transitional shock waves for generic flux functions is characterized by a *transitional map* X defined on an open set T, the transitional region. The transitional map carries each point U in T to a unique point $U' = X(U)$ in $T' = X[T]$ such that U and U' are connected by an admissible crossing discontinuity. We contrast this picture with that of Lax waves, where to each state there corresponds a curve of states. More generally, there might be several transitional maps. (For quadratic models, different maps are associated with different viscosity angles.)

Remark. We expect this map to be stable under small perturbations of the flux functions and of the viscosity matrix. Techniques such as those used in [36] should suffice to establish stability. This property is crucial to ensure that these new shock waves have physical significance.

Transitional shock waves are used to solve Riemann problems in the following manner. For a given left state U_L, the 1-wave curve based on U_L is built. If the curve

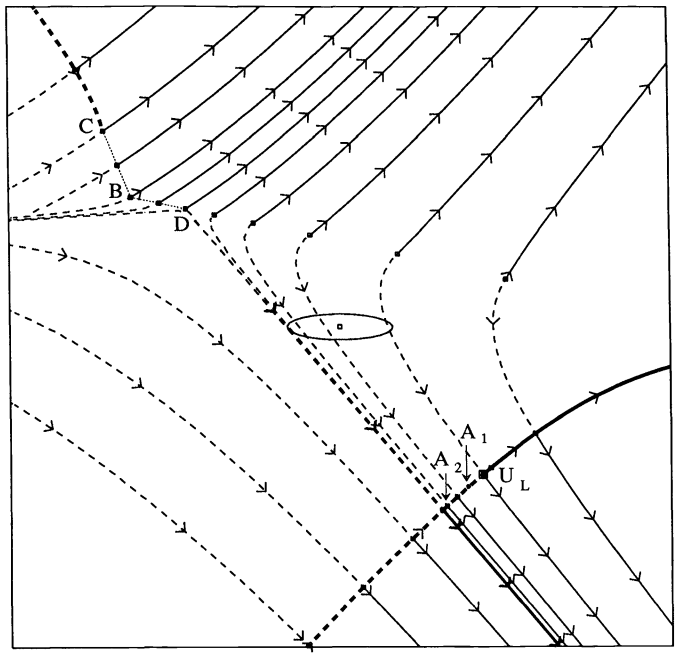


FIG. 8. *Solutions of Riemann problems for a particular left state U_L. The system is a quadratic model with an elliptic region; all shock waves admit viscous profiles for a viscosity matrix D that differs from the identity. Points between B and C correspond to transitional shock waves from points between A_2 and A_1. Points between B and D correspond to nonlocal 2-shock waves from points between A_2 and the bifurcation line, while points above C are nonlocal 1-shock waves from U_L.*

passes through the region T, then each state U_M on the curve in this region is joined to its image state $U'_M = X(U_M)$ by a transitional shock wave. Of course, the speed of this wave must exceed the speed of the 1-wave from U_L to U_M. Under this restriction, an image curve in the region T' is generated. Finally, 2-wave curves are drawn from points on this transitional curve, thus covering a region in state space. The procedure just outlined can be generalized to systems of arbitrary size.

This construction is shown in Fig. 8. (The parameter values for the model in both Figs. 5 and 8 are $a_1 = \frac{1}{2}$, $c_1 = 1 = b_2$, $e_1 = 1 = -d_2$, and all others are equal to zero. The viscosity matrix has components $D_{11} = 1.1$, $D_{12} = 0.1 = D_{21}$, and $D_{22} = 1.4$. In particular, the transitional regions are the same as in Fig. 5.) The 1-wave curve passes through region T_1, and the portion (A_1A_2) is mapped onto the curve (CB). Therefore the strip left uncovered by Lax waves is filled by solutions composed of three wave groups—a 1-wave, a transitional shock wave, and a 2-wave.

Note that because the transitional wave from A_1 to C has the same speed as the 1-shock wave from L to A_1, the points L, A_1, and C are all critical points of the dynamical system (2.10). This system has a saddle-saddle connection between A_1 and C, so that it is subject to bifurcation. Indeed, numerical evidence indicates that points on the branch of 1-shock waves above C are joined to U_L by a node-saddle connecting orbit, whereas points on the continuation of this branch below C are not, even though they correspond to Lax waves. (Observe that 1-shock waves below C are faster than at C, and that as the shock speed is increased, the saddlepoint near A_1 shifts toward U_L. When the saddle-saddle connection from A_1 to C is broken by increasing the speed, the stable manifold for the saddlepoint near A_1 shifts, blocking any connection

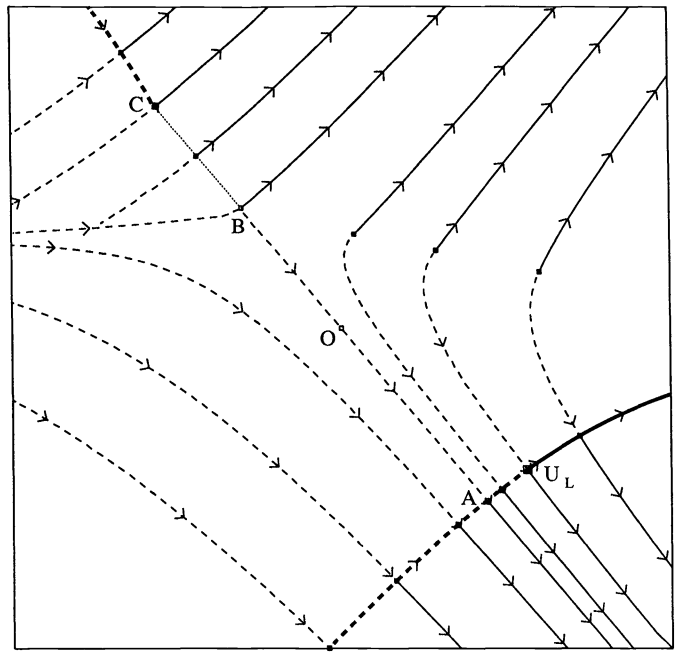


FIG. 9. *Solutions of Riemann problems for a homogeneous quadratic model in Case* II *and a particular left state* U_L. *All shock waves admit viscous profiles for* $D = I$. *Points between B and C correspond to transitional shock waves from point A.*

for nonlocal 1-shock waves.) Similarly, the transitional wave from A_2 to B is an endpoint of the transitional curve because the critical point B is not hyperbolic. Points on the curve joining points B and D are reached by (admissible) nonlocal 2-shock waves from points between A_2 and the bifurcation line.

The branches of 2-shock waves beginning at states above C end when the speed of the 2-shock wave coincides with the speed of the 1-shock wave. Such an endpoint is a totally compressive wave on the Hugoniot locus through U_L. Analogously, the branches of 2-shock waves emanating from states between C and B end when the speed of the 2-shock wave coincides with the speed of the transitional shock wave; the endpoint is a nonlocal 2-shock wave from the state between A_1 and A_2 that lies on the left side of the transitional wave. The length of the 2-shock wave branch shrinks to zero as B is approached, so that the locus of endpoints joins the totally compressive waves to state B. This locus need not coincide with the branch of nonlocal 2-shock waves from A_2 that begins at B; between it and the nonlocal branch lie admissible nonlocal 2-shock waves from states between A_1 and A_2.

The picture just described is different from that obtained in the nongeneric case of Theorem 3.5: in part (b), the transitional region is not open, being open rays of the viscosity lines, and the corresponding transitional shock waves comprise open segments on these lines. This situation arises in several examples in which transitional waves are used [35], [9], [10], [34] because D is taken to be the identity matrix. Parameter and equation counting for saddle-saddle connecting orbits are inconsistent with this picture [30]. This nongeneric case is illustrated in Fig. 9, which depicts the limits of solutions in Fig. 8 as D approaches the identity matrix and the elliptic region shrinks to zero (i.e., $e_1 = 0 = -d_2$): the curve (A_1A_2) collapses to a single point A, which is the left state for transitional shock waves to points along (CB).

Figure 9 represents part of the solution for a symmetric quadratic model in Case II [21]; this solution enforces the viscous profile admissibility criterion. In [21] and [32], solutions for Case II quadratic models are obtained using all shock waves obeying Lax's characteristic inequalities, regardless of whether these shock waves possess viscous profiles. In fact, numerical evidence suggests that nonlocal 1- and 2-shock waves that do not have viscous profiles appear in these solutions, such as nonlocal 1-shock waves on the Hugoniot branch below point C and nonlocal 2-shock waves to points near (CB) in Fig. 8. This is an instructive example in which two distinct solutions of the general Riemann problem arise from different choices of an admissibility criterion. Each of these choices yields a solution that is complete and unique.

Remark. We expect that standard numerical methods employed in solving the Cauchy problem for conservation laws are inaccurate if the solution involves transitional or nonlocal shock waves. The reasons are as follows. First, many numerical schemes spread a strong shock wave across several mesh zones, replacing it with many weak shock waves, each approximated by the local rarefaction curve; but this approximation is not valid for nonlocal shock waves, which are noncontractable. (Methods such as random choice [8] do not make this approximation.) Second, transitional shock waves are sensitive to the precise form of the diffusion term (cf. Figs. 8 and 9). In contrast to Lax shock waves, which are affected only by the overall magnitude of the viscosity, the asymptotic states in a crossing shock wave are dependent on the relative sizes of components of the viscosity matrix. Dissipative numerical schemes on coarse grids calculate transitional shock waves that correspond to the numerical viscosity, rather than the physical viscosity.

The usage of transitional rarefaction waves in solving Riemann problems is simpler; it resembles the degenerate case for transitional shock waves. As shown in § 4, transitional rarefaction curves emanate from isolated points on the coincidence locus. For instance, in Case II quadratic models there is a single transitional rarefaction curve; in Fig. 8 it passes tangent to the top of the elliptic boundary, with the 2-family portion extending to the left, and the 1-family portion extending to the right. Suppose that the 1-wave curve through U_L intersects the 2-family portion of the transitional curve at U_M; then the 1-wave from U_L to U_M can be followed by a transitional rarefaction wave from U_M to a point $U_{M'}$ on the 1-family portion, which in turn can be followed by a 2-wave. This construction is completely analogous to that used in specific examples [21], [32].

In summary, a transitional wave can appear between a 1-wave group and a 2-wave group in a solution of a Riemann problem. The procedure for constructing such solutions is to associate a transitional wave curve to the left state U_L. This transitional curve is obtained by applying the transitional map to the 1-wave curve in the case of shock waves, and by following an integral curve in the case of rarefaction waves. More generally, composite waves containing transitional waves along with Lax waves can occur [17], [14] (see Fig. 2c, d).

6. Summary. Transitional waves, which are not associated with a particular characteristic family, arise in non-strictly hyperbolic systems of conservation laws. Because of such waves, the solution of a Riemann problem for a system of n conservation laws might contain more than n wave groups.

Transitional shock waves are discontinuous solutions that possess viscous profiles corresponding to saddle-saddle connecting orbits. For transitional shock waves, the association of U_L with U_R is a map defined on a region in state space (for generic viscosity matrices); this has been demonstrated explicitly for conservation laws with

quadratic flux functions, where saddle-saddle connecting orbits are straight-line segments.

Transitional rarefaction waves are continuous solutions that switch from a faster family to a slower one. Using a geometric framework, the generic nature of rarefaction waves near an elliptic region in systems of two conservation laws has been established: if a rarefaction curve intersects the elliptic boundary, then, except at isolated points, it switches from a slower family to a faster one and forms a cusp; the exceptional points are where the rarefaction curve is transitional and passes tangent to the elliptic region.

Transitional waves play a significant role in solving the Riemann problem for non-strictly hyperbolic systems. This is illustrated in a quadratic model for which the general Riemann problem has two distinct solutions, both complete and unique, depending on the admissibility criterion imposed on shock waves. One solution uses all waves satisfying the characteristic criterion, some of which do not possess viscous profiles; the other uses the viscous profile criterion and requires transitional shock waves.

Appendix A. Proof of Corollary 3.9. In this Appendix, we establish the nature of critical points at infinity for quadratic conservation laws.

Consider a quadratic dynamical system in the plane

$$\dot{u} = P(u, v), \tag{A1}$$

$$\dot{v} = Q(u, v), \tag{A2}$$

where P and Q are quadratic polynomials in u and v and the dot denotes differentiation with respect to the independent variable ξ. The behavior of solutions is affected not only by critical points (u_c, v_c) in the finite plane, where $P(u_c, v_c) = 0$ and $Q(u_c, v_c) = 0$, but also by the behavior of P and Q near infinity, i.e., the asymptotic directions of the vector field.

To study the behavior near infinity, we make a (singular) change of independent variables from ξ to η by setting $d/d\eta = R^{-1} d/d\xi$, where $R = (u^2 + v^2)^{1/2}$. This allows us to exploit the approximate homogeneity of P and Q for large R:

$$P(u, v) = R^2\{P_2(u/R, v/R) + O(R^{-1})\}, \tag{A3}$$

$$Q(u, v) = R^2\{Q_2(u/R, v/R) + O(R^{-1})\}, \tag{A4}$$

P_2 and Q_2 being the homogeneous quadratic parts of P and Q, respectively. When we introduce polar coordinates, $u = R\cos\varphi$ and $v = R\sin\varphi$, and denoting $\rho = R^{-1}$, a straightforward calculation shows that

$$\varphi' = \cos\varphi\, Q_2(\cos\varphi, \sin\varphi) - \sin\varphi\, P_2(\cos\varphi, \sin\varphi) + O(\rho), \tag{A5}$$

$$\rho' = -\rho\{\cos\varphi\, P_2(\cos\varphi, \sin\varphi) + \sin\varphi\, Q_2(\cos\varphi, \sin\varphi) + O(\rho)\}, \tag{A6}$$

where the prime denotes differentiation with respect to η. Therefore a *critical point at infinity* occurs if and only if $\rho = 0$ and

$$\cos\varphi\, Q_2(\cos\varphi, \sin\varphi) - \sin\varphi\, P_2(\cos\varphi, \sin\varphi) = 0. \tag{A7}$$

Equation (A7) is a homogeneous cubic polynomial in $\sin\varphi$ and $\cos\varphi$; its roots give the asymptotic directions of the vector field. The eigenvalues of the linearization of

(A5) and (A6) near a critical point at infinity are

$$\mu_\varphi = \frac{d}{d\varphi}\{\cos\varphi\, Q_2(\cos\varphi, \sin\varphi) - \sin\varphi\, P_2(\cos\varphi, \sin\varphi)\}, \tag{A8}$$

$$\mu_\rho = -\{\cos\varphi\, P_2(\cos\varphi, \sin\varphi) + \sin\varphi\, Q_2(\cos\varphi, \sin\varphi)\}, \tag{A9}$$

corresponding to the φ and ρ directions.

For the dynamical system

$$D\dot{U}(\xi) = -s[U(\xi) - U_-] + F(U(\xi)) - F(U_-) \tag{A10}$$

derived from a quadratic system of conservation laws, $(P_2, Q_2)^T = D^{-1}F_2$, where F_2 is the homogeneous quadratic part of the flux F. Let α_D, β_D, $\tilde{\alpha}_D$, and $\tilde{\beta}_D$ denote the functions associated with $D^{-1}F$ that are defined (for F) in (3.7), (3.8), (3.10), and (3.11). In these terms, a critical point at infinity occurs precisely when $\rho = 0$ and

$$\alpha_D(\varphi)\cos\varphi + \beta_D(\varphi)\sin\varphi = 0, \tag{A11}$$

and the eigenvalues at such a point are

$$\mu_\varphi = \frac{1}{2}\frac{d}{d\varphi}\{\alpha_D(\varphi)\cos\varphi + \beta_D(\varphi)\sin\varphi\}, \tag{A12}$$

$$\mu_\rho = -\frac{1}{2}\{\tilde{\alpha}_D(\varphi)\cos\varphi + \tilde{\beta}_D(\varphi)\sin\varphi\}. \tag{A13}$$

By definition, an angle φ satisfying (A11) is a viscosity angle; thus critical points at infinity occur at viscosity angles. Furthermore, $\mu_\varphi \neq 0$ for a critical point at infinity if the corresponding root of (A11) is simple, i.e., the viscosity angle φ is simple, while $\mu_\rho \neq 0$ if φ is not exceptional. If $\mu_\varphi \mu_\rho \neq 0$, then the topological index of the critical point at infinity is $\mathrm{Ind}(\varphi) = \mathrm{sgn}(\mu_\varphi \mu_\rho)$. Consequently, we have demonstrated the following.

LEMMA A1. *Consider a quadratic model together with a viscosity matrix D. Then critical points at infinity for the dynamical system* (A10) *occur precisely at viscosity angles. If the viscosity angle φ is simple and nonexceptional, then the topological index of the corresponding critical point at infinity is*

$$\mathrm{Ind}(\varphi) = -\mathrm{sgn}\left[\frac{d}{d\varphi}\{\alpha_D(\varphi)\cos\varphi + \beta_D(\varphi)\sin\varphi\}\{\tilde{\alpha}_D(\varphi)\cos\varphi + \tilde{\beta}_D(\varphi)\sin\varphi\}\right]. \tag{A14}$$

Appendix B. Examples of characteristic surfaces. In this Appendix, we present two examples of the constructions of § 4 for rarefaction waves.

B.1. Quadratic models. The approach of § 4 was developed by Palmeira [28] for quadratic models with compact elliptic regions; the results carry over to general quadratic models. Using the notation of § 3.1, it is simple to verify that

$$\mathscr{F}(u, v, \varphi) = \alpha(\varphi)u + \beta(\varphi)v + \gamma(\varphi), \tag{B1}$$

$$\lambda(u, v, \varphi) = \tilde{\alpha}(\varphi)u + \tilde{\beta}(\varphi)v + \tilde{\gamma}(\varphi). \tag{B2}$$

Thus the characteristic surface $\mathscr{F} = 0$ is ruled: each horizontal plane $\varphi = \text{const.}$ intersects the surface in a straight line $\alpha(\varphi)u + \beta(\varphi)v + \gamma(\varphi) = 0$, the projection of which is a characteristic line. The surface is regular except when $\alpha(\varphi) = \beta(\varphi) = \gamma(\varphi) = 0$ and $\mathscr{F}_\varphi = 0$ (see (B4) below).

The coincidence locus is defined by the equations $\mathscr{F} = 0$ and $\mathscr{F}_\varphi = 0$, i.e.,

$$\alpha(\varphi)u + \beta(\varphi)v + \gamma(\varphi) = 0, \tag{B3}$$

$$\alpha'(\varphi)u + \beta'(\varphi)v + \gamma'(\varphi) = 0. \tag{B4}$$

Note that these are also the equations for the envelope of the characteristic lines. For simplicity, we assume that the determinant $\mathscr{D} = \alpha\beta' - \beta\alpha'$ does not vanish identically. Then the linear equations (B3) and (B4) may be solved to express u and v on the coincidence locus in terms of φ: $u = u_c(\varphi)$ and $v = v_c(\varphi)$, where

$$u_c = -[\gamma\beta' - \beta\gamma']/\mathscr{D}, \tag{B5}$$

$$v_c = -[\alpha\gamma' - \gamma\alpha']/\mathscr{D}. \tag{B6}$$

Despite appearances, $\mathscr{D}$ and the numerators of u_c and v_c are linear, not quadratic, in $\sin 2\varphi$ and $\cos 2\varphi$. Thus the coincidence locus is a conic section; any asymptotes occur at the angles φ where $\mathscr{D}(\varphi) = 0$.

A coincidence point is a foldpoint unless $\mathscr{F}_{\varphi\varphi} = 0$. Evaluated on the surface $\mathscr{F} = 0$,

$$\tfrac{1}{4}\mathscr{F}_{\varphi\varphi}(u, v, \varphi) = \alpha_0 u + \beta_0 v + \gamma_0,$$

where $\alpha_0 = \frac{1}{2}(a_2 - b_1)$, $\beta_0 = \frac{1}{2}(b_2 - c_1)$, and $\gamma_0 = \frac{1}{2}(d_2 - e_1)$ are the φ-independent parts of α, β, and γ. Therefore a coincidence point is a foldpoint so long as $\alpha_0 u_c(\varphi) + \beta_0 v_c(\varphi) + \gamma_0 \neq 0$. A simplification occurs here also: $\alpha_0 u_c + \beta_0 v_c + \gamma_0 = -\frac{1}{4}\eta/\mathscr{D}$ with η constant. Provided that $\eta \neq 0$, all points on the coincidence locus are foldpoints.

In terms of the parameterization of the coincidence locus, any solution (u, v) of $\mathscr{F}(u, v, \varphi) = 0$ takes the form

$$u = u_c(\varphi) - \beta(\varphi) \cdot \kappa, \tag{B7}$$

$$v = v_c(\varphi) + \alpha(\varphi) \cdot \kappa \tag{B8}$$

for some $\kappa \in \mathbf{R}$ (except when $\mathscr{D}(\varphi) = 0$). Thus φ and κ give global coordinates for the characteristic surface, and $\kappa = 0$ defines the coincidence locus.

To determine the differential equations for integral curves in these coordinates, we require formulae for the derivatives u_c' and v_c'. Substituting u_c and v_c into (B3) and (B4) and differentiating shows that

$$\alpha u_c' + \beta v_c' = 0, \tag{B9}$$

$$\alpha' u_c' + \beta' v_c' = -4[\alpha_0 u_c + \beta_0 v_c + \gamma_0]. \tag{B10}$$

Thus $u_c' = -\beta\eta/\mathscr{D}^2$ and $v_c' = \alpha\eta/\mathscr{D}^2$.

Expressed in global coordinates, the equation defining integral curves is

$$\begin{aligned} 0 &= -\sin\varphi\, du + \cos\varphi\, dv \\ &= [\alpha\cos\varphi + \beta\sin\varphi]\, d\kappa \\ &\quad + \{[\alpha\cos\varphi + \beta\sin\varphi]\eta/\mathscr{D}^2 + [\alpha'\cos\varphi + \beta'\sin\varphi]\cdot\kappa\}\, d\varphi. \end{aligned} \tag{B11}$$

Thus integral curves may be obtained (locally) by solving

$$\dot{\varphi} = -[\alpha\cos\varphi + \beta\sin\varphi], \tag{B12}$$

$$\dot{\kappa} = [\alpha\cos\varphi + \beta\sin\varphi]\eta/\mathscr{D}^2 + [\alpha'\cos\varphi + \beta'\sin\varphi]\cdot\kappa. \tag{B13}$$

These equations yield a first-order, linear differential equation for κ as a function of φ. Critical points of the dynamical system (B12) and (B13) occur precisely when $\alpha\cos\varphi + \beta\sin\varphi = 0$ and $\kappa = 0$, i.e., at coincidence points for which φ is an asymptotic angle. In other words, the critical points occur where the bifurcation lines are tangent to the coincidence locus.

The inflection locus is defined by (4.18). To solve this equation in the present case, note that $\lambda_\varphi = -\frac{1}{2}\mathscr{F}_{\varphi\varphi} = \frac{1}{2}[\eta/\mathscr{D} - \mathscr{D}' \cdot \kappa]$ and that $\mathscr{F}_\varphi = \mathscr{D} \cdot \kappa$. Then the inflection locus is determined by the equation

(B14) $$[(\tfrac{1}{2}\mathscr{D}'\alpha + \mathscr{D}\tilde{\alpha}) \cos\varphi + (\tfrac{1}{2}\mathscr{D}'\beta + \mathscr{D}\tilde{\beta}) \sin\varphi] \cdot \kappa = \tfrac{1}{2}[\alpha \cos\varphi + \beta \sin\varphi]\eta/\mathscr{D}.$$

The quantities in parentheses in this equation simplify to linear expressions in $\sin 2\varphi$ and $\cos 2\varphi$.

B.2. Keyfitz–Kranzer models. Systems of conservation laws of the form

(B15) $$u_t + [u\Phi(u, v)]_x = 0,$$

(B16) $$v_t + [v\Phi(u, v)]_x = 0$$

have been studied as models for elastic strings [23], [24] and for multiphase flows in petroleum reservoirs [16], [37], [22], [27]. For such a system,

(B17) $$\mathscr{F}(u, v, \varphi) = [-u \sin\varphi + v \cos\varphi][\Phi_u(u, v) \cos\varphi + \Phi_v(u, v) \sin\varphi],$$

(B18) $$\lambda(u, v, \varphi) = \Phi(u, v) + [u \cos\varphi + v \sin\varphi][\Phi_u(u, v) \cos\varphi + \Phi_v(u, v) \sin\varphi].$$

Therefore the characteristic surface is the intersection of two surfaces: the ruled surface where $\mathscr{F}_{\text{ruled}}(u, v, \varphi) = -u \sin\varphi + v \cos\varphi$ is zero; and the surface where $\mathscr{F}_{\text{contact}}(u, v, \varphi) = \Phi_u(u, v) \cos\varphi + \Phi_v(u, v) \sin\varphi$ is zero, which we call the "contact" surface because of its relation to the linearly degenerate wave mode.

For a point (u, v, φ) on the ruled surface, $u = \kappa \cos\varphi$ and $v = \kappa \sin\varphi$ for some $\kappa \in \mathbf{R}$, so that $\mathscr{F}_\varphi(u, v, \varphi) = -\kappa\mathscr{F}_{\text{contact}}(u, v, \varphi)$. In particular, $\lambda = \Phi + u\Phi_u + v\Phi_v$. Similarly, for a point on the contact surface, $\Phi_u = -\mu \sin\varphi$ and $\Phi_v = \mu \cos\varphi$ for some $\mu \in \mathbf{R}$, $\mathscr{F}_\varphi(u, v, \varphi) = \mu\mathscr{F}_{\text{ruled}}(u, v, \varphi)$, and $\lambda = \Phi$.

It follows that the coincidence locus, where both $\mathscr{F}$ and $\mathscr{F}_\varphi$ vanish, comprises the intersection of the ruled and contact surfaces, together with the vertical line $u = 0$, $v = 0$. Necessarily, the characteristic surface fails to be regular at coincidence points, where it is not a manifold; indeed, $\mathscr{F}_u$ and $\mathscr{F}_v$, as well as $\mathscr{F}_\varphi$, vanish on the coincidence locus. The projection of this locus onto the u-v plane is given by the equation

(B19) $$u\Phi_u + v\Phi_v = 0,$$

and it is bordered on both sides by regions of strict hyperbolicity, instead of separating a hyperbolic region from an elliptic region.

Remark. In one sense, the system of conservation laws (B16) and (B17) is not generic: the flux functions take a special form $f(u, v) = u\Phi(u, v)$ and $g(u, v) = v\Phi(u, v)$. This form is not stable under general perturbations, which would break the intersecting surfaces apart and remove the singularity. However, the conservation laws may be regarded as generic, in a different sense, provided that the physical system being modeled imposes this special form of the flux functions and allows for general perturbations of Φ.

Evaluated on the ruled surface, the equation defining integral curves is

(B20) $$0 = -\sin\varphi\, du + \cos\varphi\, dv = \kappa\, d\varphi.$$

Thus an integral curve in this part of the characteristic surface is a horizontal line $\varphi = \text{const.}$, $-u \sin\varphi + v \cos\varphi = 0$. On the contact surface,

(B21) $$0 = \mu[-\sin\varphi\, du + \cos\varphi\, dv] = d\Phi,$$

which implies that $\Phi = \text{const.}$ along integral curves in this portion. All points on the coincidence locus are critical points for the line field defining integral curves, so that Theorem 4.2 does not apply.

It may be verified that the determinant of (4.18), which defines the inflection locus, vanishes identically on the contact surface; therefore the eigenvalue corresponding to this part of the characteristic surface is linearly degenerate. On the ruled surface, the equation for the inflection locus reduces to

$$u[\Phi + u\Phi_u + v\Phi_v]_u + v[\Phi + u\Phi_u + v\Phi_v]_v = 0, \tag{B22}$$

i.e., to the vanishing of the derivative of the eigenvalue along the rarefaction.

In summary, the behavior of rarefactions near the coincidence locus, which separates two regions of strict hyperbolicity, is as follows. A rarefaction curve projected from the contact surface is a level curve $\Phi = \text{const.}$, whereas a rarefaction curve projected from the ruled surface lies along a line $-u \sin \varphi + v \cos \varphi = 0$, $\varphi = \text{const.}$ The two types of rarefaction curves are tangent to each other at coincidence points, and generically they cross the coincidence locus transversally. The corresponding eigenvalues are $\lambda_{\text{contact}} = \Phi$ and $\lambda_{\text{ruled}} = \Phi + u\Phi_u + v\Phi_v$, respectively. Thus the contact curves are linearly degenerate, while the other eigenvalue typically increases monotonically through the coincidence locus. In particular, λ_{ruled} switches from family 1 to family 2 as the corresponding rarefaction curve is followed, in the direction of increasing eigenvalue, across the coincidence locus; at the same time, λ_{contact} switches from family 2 to family 1.

Acknowledgments. We thank Professor Mark Ashbaugh for bringing the work of Chicone to our attention. We are also grateful to Professor Geovan Tavares dos Santos and Dr. M. Elasir Gomes for discussions concerning Theorem 3.8, and to Professor C. Frederico Palmeira for conversations about transitional rarefaction waves.

The hospitality of the Courant Institute of Mathematical Sciences of New York University, the Departments of Mathematics at Pontifícia Universidade Católica of Rio de Janeiro and at the University of Wyoming, the Department of Computer Sciences and the Mathematics Research Center at the University of Wisconsin, Madison, and the Instituto de Matemática Pura e Aplicada is gratefully acknowledged.

REFERENCES

[1] A. AZEVEDO, *Multiple viscous profile solutions of the Riemann problem for a mixed elliptic-hyperbolic system*, Ph.D. thesis, Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro, in preparation.

[2] C. CHICONE, *Quadratic gradients on the plane are generically Morse-Smale*, J. Differential Equations, 33 (1979), pp. 159-166.

[3] C. CHICONE AND T. JINGHUANG, *On general properties of quadratic systems*, Amer. Math. Monthly, 89 (1982), pp. 167-178.

[4] C. CONLEY AND J. SMOLLER, *Viscosity matrices for two-dimensional nonlinear hyperbolic systems*, Comm. Pure Appl. Math., 23 (1970), pp. 867-884.

[5] R. COURANT AND K. FRIEDRICHS, *Supersonic Flow and Shock Waves*, John Wiley, New York, 1948.

[6] L. FOY, *Steady state solutions of hyperbolic systems of conservation laws with viscous terms*, Comm. Pure Appl. Math., 17 (1964), pp. 177-188.

[7] I. GELFAND, *Some problems in the theory of quasi-linear equations*, Uspekhi Mat. Nauk, 14 (1959), pp. 87-158. (In Russian.) Amer. Math. Soc. Transl., Ser. 2, 29 (1963), pp. 295-381. (In English.)

[8] J. GLIMM, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697-715.

[9] M. E. GOMES, *Singular Riemann problem for a fourth-order model for multi-phase flow*, Ph.D. thesis, Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro, 1987. (In Portuguese.)

[10] ———, *Riemann problems requiring a viscous profile entropy condition*, Adv. in Appl. Math., 10 (1989), pp. 285-323.

[11] ———, *On saddle connections of quadratic dynamical systems with application to conservation laws*, preprint, Courant Institute of the Mathematical Sciences, New York University, New York, 1989.

[12] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, Berlin, New York, 1986.

[13] H. HOLDEN, *On the Riemann problem for a prototype of a mixed type conservation law*, Comm. Pure Appl. Math., 40 (1987), pp. 229-264.

[14] H. HOLDEN AND L. HOLDEN, *On the Riemann problem for a prototype of a mixed type conservation law*, II, preprint, University of Trondheim, Trondheim, Norway, 1988.

[15] E. HOPF, *The partial differential equation* $u_t + uu_x = \mu u_{xx}$, Comm. Pure Appl. Math., 3 (1950), pp. 201-230.

[16] E. ISAACSON, *Global solution of a Riemann problem for a nonstrictly hyperbolic system of conservation laws arising in enhanced oil recovery*, J. Comp. Phys., to appear.

[17] E. ISAACSON, D. MARCHESIN, B. PLOHR, AND J. B. TEMPLE, *Multiphase flow models with singular Riemann problems*, preprint, University of Wyoming, Laramie, WY, 1988.

[18] ———, *The Riemann problem near a hyperbolic singularity: the classification of quadratic Riemann problems* I, SIAM J. Appl. Math., 48 (1988), pp. 1009-1032.

[19] E. ISAACSON AND J. B. TEMPLE, *Examples and classification of non-strictly hyperbolic systems of conservation laws*, Abstracts Amer. Math. Soc., 1985.

[20] ———, *The Riemann problem near a hyperbolic singularity* II, SIAM J. Appl. Math., 48 (1988), pp. 1287-1301.

[21] ———, *The Riemann problem near a hyperbolic singularity* III, SIAM J. Appl. Math., 48 (1988), pp. 1302-1312.

[22] ———, *The structure of asymptotic states in a singular system of conservation laws*, Adv. in Appl. Math., to appear.

[23] B. KEYFITZ AND H. KRANZER, *A system of non-strictly hyperbolic conservation laws arising in elasticity theory*, Arch. Rational Mech. Anal., 72 (1980), pp. 219-241.

[24] ———, *The Riemann problem for a class of hyperbolic conservation laws exhibiting a parabolic degeneracy*, J. Differential Equations, 47 (1983), pp. 35-65.

[25] P. LAX, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537-566.

[26] T.-P. LIU, *The Riemann problem for general* 2×2 *conservation laws*, Trans. Amer. Math. Soc., 199 (1974), pp. 89-112.

[27] J. DA MOTA, *Fundamental solutions for thermal flow of multiphase fluids in porous media*, Ph.D thesis, Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro, 1988. (In Portuguese.)

[28] C. F. PALMEIRA, *Line fields defined by eigenspaces of derivatives of maps from the plane to itself*, preprint, Pontifícia Universidade Católica do Rio de Janeiro, 1987; Proc. Sixth International Conference on Differential Geometry, Santiago de Compostela, Spain, 1988.

[29] M. RASCLE, *The Riemann problem for a nonlinear non-strictly hyperbolic system arising in biology*, preprint, Université de Saint-Etienne, France, 1986.

[30] D. SCHAEFFER AND M. SHEARER, private communication, 1985.

[31] ———, *The classification of* 2×2 *systems of non-strictly hyperbolic conservation laws, with application to oil recovery*, Comm. Pure Appl. Math., 40 (1987), pp. 141-178.

[32] ———, *Riemann problems for nonstrictly hyperbolic* 2×2 *systems of conservation laws*, Trans. Amer. Math. Soc., 304 (1987), pp. 267-306.

[33] D. SERRE, *Existence globale de solutions faibles sous une hypothèse unilatérale*, Quart. J. Appl. Math., 46 (1988), pp. 157-167.

[34] M. SHEARER, *The Riemann problem for* 2×2 *systems of hyperbolic conservation laws with case* I *quadratic nonlinearities*, preprint, North Carolina State University, Raleigh, NC, 1988.

[35] M. SHEARER, D. SCHAEFFER, D. MARCHESIN, AND P. PAES-LEME, *Solution of the Riemann problem for a prototype* 2×2 *system of non-strictly hyperbolic conservation laws*, Arch. Rational Mech. Anal., 97 (1987), pp. 299-320.

[36] M. SHEARER AND S. SCHECTER, *Riemann problems involving undercompressive shocks*, in Workshop on Partial Differential Equations and Continuum Models of Phase Transitions (Nice, 1988), D. Serre, ed., Lecture Notes in Math., Springer-Verlag, New York, 1988.

[37] J. B. TEMPLE, *Global solution of the Cauchy problem for a class of* 2×2 *non-strictly hyperbolic conservation laws*, Adv. in Appl. Math., 3 (1982), pp. 335-375.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 867-883, July 1990
© 1990 Society for Industrial and Applied Mathematics
003

ASYMPTOTIC ANALYSIS ON LARGE TIMESCALES FOR SINGULAR PERTURBATIONS OF HYPERBOLIC TYPE*

W. ECKHAUS† AND M. GARBEY‡

Abstract. A uniform approximation of a linear hyperbolic-hyperbolic singular perturbation problem for a large timescale under a "timelike" hypothesis is constructed. It is shown that the effect of a hyperbolic perturbation is qualitatively the same as the effect of a viscous perturbation for a large timescale. The validity is proved by the use of the energy method.

Key words. singular perturbations, hyperbolic equations, asymptotic analysis, large timescales

AMS(MOS) subject classification. 65

1. Introduction.

We consider the following singular perturbation problem:

$$P_\varepsilon \begin{cases} L_\varepsilon[u] = \varepsilon L_2[u] + L_1[u] = f(x, t, \varepsilon), \\ u(x, 0, \varepsilon) = g(x), \\ \dfrac{\partial u}{\partial t}(x, 0, \varepsilon) = h(x), \end{cases} \tag{1.1}$$

where x is the space variable, $x \in R$, t represents time, $t \geqq 0$, and ε is a small positive parameter. L_2 is a strictly hyperbolic linear operator of second order:

$$L_2[u] = \left(\frac{\partial}{\partial t} + c_1(x, t)\frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial t} + c_2(x, t)\frac{\partial}{\partial x}\right)[u]$$

with $c_1 < c_2$, and L_1 is a nondegenerate hyperbolic linear operator of first order:

$$L_1[u] = a(x, t, \varepsilon)\frac{\partial u}{\partial t} + b(x, t, \varepsilon)\frac{\partial u}{\partial x} + d(x, t, \varepsilon) \cdot u$$

with $a(x, t, \varepsilon) \geqq a_0 > 0$. We suppose "time behavior" for L_ε, and to apply the theory of asymptotic developments, we need suitable regularity for the given functions f, g, h, a, b, c_i, d. This type of problem occurs in many physical models (see Whitham [6, Chaps. 3 and 10, "Wave Hierarchies"]).

The initial boundary layer has been studied by several authors, in particular, Geel and Dejager (see [3] and its bibliography) and Genet and Madaune-Tort (see [5] and its bibliography and [8]). It has been established (see [3], [4]) that for time of order 1 (strictly), the solution of the reduced problem (formal limit of P_ε):

$$P_0\colon L_1[w_0] = f(x, t, 0), \qquad w_0(x, 0) = g(x) \tag{1.2}$$

is a good approximation of u, the solution of P_ε (outside the initial boundary layer), i.e.,

$$\begin{cases} u = w_0 + O_s(\varepsilon), \\ u_x = \dfrac{\partial w_0}{\partial x} + O_s(\varepsilon), \qquad t \in [p, T], \quad 0 < p < T, \\ u_t = \dfrac{\partial w_0}{\partial t} + O_s(\varepsilon), \end{cases}$$

* Received by the editors June 26, 1988; accepted for publication (in revised form) July 25, 1989.
† Mathematisch Instituut, Utrecht, the Netherlands.
‡ Université de Valenciennes et du Hainaut-Cambresis, Valenciennes, France.

where p is an arbitrary constant. The definition of the O_s symbol is as in Eckhaus [1]: $u - w = O_s(\varepsilon)$ means $u - w = O(\varepsilon)$ and $u - w \neq o(\varepsilon)$. In this paper we are concerned with the behavior of u for large timescales, namely, $1/t = o(1)$. The contribution of the perturbation cannot be neglected in this case, and the wave w_0 does not describe (even in first approximation) the evolution of u.

Let us explain the salient feature of our method, in the case of constant coefficients, more precisely for the equation

$$\varepsilon\left(\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2}\right) + \frac{\partial u}{\partial t} + b\frac{\partial u}{\partial x} = 0 \quad \text{with } |b| < 1. \tag{1.3}$$

The last condition assures "timelike" behavior. See Fig. 1. Following Eckhaus [2], we note that the solutions of the unperturbed equation $\partial u/\partial t + b(\partial u/\partial x) = 0$, represent "waves" that move with a constant speed b without changing shape. To extend the observation over a large period of time, we follow these waves using the characteristic transformation $\tilde{x} = x - bt$. Let $\bar{u}(\tilde{x}, t, \varepsilon) = u(x, t, \varepsilon)$. We rewrite (1.3) as

$$\varepsilon\left(-(1-b^2)\frac{\partial^2 \bar{u}}{\partial \tilde{x}^2} - 2b\frac{\partial^2 \bar{u}}{\partial \tilde{x}\,\partial t} + \frac{\partial^2 \bar{u}}{\partial t^2}\right) + \frac{\partial \bar{u}}{\partial t} = 0. \tag{1.4}$$

Now there are only two significant scales of time:

(i) $\hat{t} = t/\varepsilon$, $(\hat{u}(\tilde{x}, \hat{t}, \varepsilon) = u(x, t, \varepsilon))$.

Then formally

$$\frac{\partial^2 \hat{u}}{\partial \hat{t}^2} + \frac{\partial \hat{u}}{\partial \hat{t}} = O(\varepsilon),$$

which describes the initial layer already studied in Geel [3].

(ii) $\tau = t\varepsilon$, $(\tilde{u}(\tilde{x}, \tau, \varepsilon) = u(x, t, \varepsilon))$.

Then formally $\partial \tilde{u}/\partial \tau = (1-b^2)\partial^2 \tilde{u}/\partial \tilde{x}^2 + O(\varepsilon)$, which describes the behavior for a large timescale. Note that the formal limit

$$\frac{\partial \tilde{u}_0}{\partial \tau} = (1-b^2)\frac{\partial^2 \tilde{u}_0}{\partial \tilde{x}^2}$$

is a forward diffusion equation if the subcharacteristics of (1.3) are "timelike." It may seem surprising at first sight that solutions of a hyperbolic equation (i.e., with a finite propagation velocity) are governed for long times by a parabolic equation. In fact, we

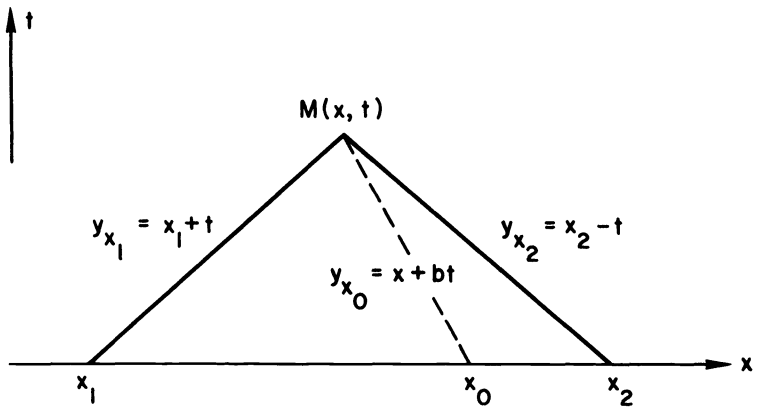


FIG. 1. *Timelike subcharacteristics.*

will show (as a by-product of our general analysis) that in the case of equation (1.3) the solution of the reduced problem

$$\frac{\partial \tilde{u}_0}{\partial \tau} = (1-b^2)\frac{\partial^2 \tilde{u}_0}{\partial \tilde{x}^2}, \qquad \tilde{u}_0(\tilde{x}, 0) = g(\tilde{x})$$

is a good approximation of u (outside the initial boundary layer) that contains the approximation w_0 for t strictly of order 1 and that is still valid for t of order $1/\varepsilon$; more precisely, we have that $u = \tilde{u}_0 + O_s(\varepsilon)$ for $t \in [p, T/\varepsilon]$; $p, T > 0$; p, T arbitrary constants.

The main body of our paper is devoted to asymptotic analysis of the general problem P_ε, with some suitable additional hypotheses on the behavior of the coefficients.

We will show that in the general case the perturbation also induces a diffusion process similar to that for the simple case outlined above.

In the following we do the asymptotic analysis of P_ε for large timescales in two steps: first, we formally do the construction of the development to any order ε^m, $m \in N$, of the solution $u(x, t, \varepsilon)$ of P_ε and natural restrictive hypotheses appear on the behavior of d, f, and subcharacteristics of P_ε for large timescales. Second, by using the energy method, we prove the correctness of the formal approximation with estimates for u and its first derivatives in the maximum norm.

2. Asymptotic analysis. Construction of a formal approximation. We look for an asymptotic approximation of the solution $u(x, t, \varepsilon)$ of the following initial value problem:

$$\varepsilon\left[\frac{\partial^2 u}{\partial t^2} + ((c_1 x, t) + c_2(x, t))\frac{\partial^2 u}{\partial x\, \partial t} + c_1(x, t)c_2(x, t)\frac{\partial^2 u}{\partial x^2}\right]$$

$$+ a(x, t, \varepsilon)\frac{\partial u}{\partial t} + b(x, t, \varepsilon)\frac{\partial u}{\partial x} + d(x, t, \varepsilon)u = f(x, t, \varepsilon), \qquad x \in R, \quad t \geqq 0, \tag{2.1}$$

$$u(x, 0, \varepsilon) = g(x), \quad \frac{\partial u}{\partial t}(x, 0, \varepsilon) = h(x), \quad x \in R$$

where

$$a(x, t, \varepsilon) \geqq a_0 > 0, \quad \text{for } x \in R,\ t \geqq 0, \text{ and } a_0 \text{ a constant independent of } \varepsilon. \tag{2.2}$$

The "timelike" hypothesis is expressed by

$$p + c_1(x, t) < b(x, t, \varepsilon)/a(x, t, \varepsilon) < c_2(x, t) - p, \quad \text{for } x \in R, \quad t \geqq 0, \quad p > 0 \text{ independent of } \varepsilon. \tag{2.3}$$

The given functions a, b, d, c_i, f, g, h are sufficiently smooth functions (the required order of differentiability depends on the order of asymptotic approximation we wish to obtain) that are uniformly bounded for $x \in R$, $t \geqq 0$.

To simplify the presentation, we will assume $a(x, t)$ and $b(x, t)$ are independent of ε. We introduce the transformation

$$(x, t) \to (\tilde{x}(x, t), \tau = t\varepsilon). \tag{2.4}$$

The function $\tilde{x}$ has yet to be determined. Let us put

$$\alpha(x, t) = \frac{\partial \tilde{x}}{\partial t}, \qquad \beta(x, t) = \frac{\partial \tilde{x}}{\partial x}.$$

We rewrite (1.1) in new variables ($\tilde{u}(\tilde{x}, \tau, \varepsilon) = u(x, t, \varepsilon)$):

$$\varepsilon(\alpha + c_1\beta)(\alpha + c_2\beta)\frac{\partial^2 \tilde{u}}{\partial \tilde{x}^2} + \varepsilon L_2[\tilde{x}]\frac{\partial \tilde{u}}{\partial \tilde{x}} + \varepsilon a \frac{\partial \tilde{u}}{\partial \tau}$$
$$+ \varepsilon^2(2\alpha + (c_1 + c_2)\beta)\frac{\partial^2 \tilde{u}}{\partial \tilde{x}\,\partial \tau} + \varepsilon^3 \frac{\partial^2 \tilde{u}}{\partial \tau^2} + (a\alpha + b\beta)\frac{\partial \tilde{u}}{\partial \tilde{x}} + d\tilde{u} - f = 0. \tag{2.5}$$

Now, we make the following remark. We know from [3] that w_0, the solution of the reduced problem

$$a(x, t)\frac{\partial w_0}{\partial t} + b(x, t)\frac{\partial w_0}{\partial x} + d(x, t, \varepsilon)w_0 = f(x, t, \varepsilon),$$
$$w_0(x, 0) = g(x), \tag{2.6}$$

is a valid approximation of u for t of order 1, namely, $u = w_0 + O_s(\varepsilon)$ for $t \in [0, T]$, T an arbitrary constant.

From Kaplun's extension lemma, there exists an order function $\delta(\varepsilon) = o(1)$ such that w_0 is still a valid approximation of u for t of order $T/\delta(\varepsilon)$. Then asymptotic analysis for large timescales is interesting if w_0 does not explode or vanish exponentially. For this purpose, we will assume

$$f(x, t, \varepsilon) = \varepsilon f_1(x, t, \varepsilon), \qquad d(x, t, \varepsilon) = \varepsilon d_1(x, t, \varepsilon)$$

with f_1, d_1 uniformly bounded for $x \in R$, $t \geqq q$, q an arbitrary constant independent of ε.

We rewrite (2.5) as follows:

$$(\alpha + c_1\beta)(\alpha + c_2\beta)\frac{\partial^2 \tilde{u}}{\partial \tilde{x}^2} + L_2[\tilde{x}]\frac{\partial \tilde{u}}{\partial \tilde{x}} + d_1\tilde{u} - f_1 + a\frac{\partial \tilde{u}}{\partial \tau}$$
$$+ \varepsilon(2\alpha + (c_1 + c_2)\beta)\frac{\partial^2 \tilde{u}}{\partial \tilde{x}\,\partial \tau} + \varepsilon^2 \frac{\partial^2 \tilde{u}}{\partial \tau^2} + \varepsilon^{-1}(a\alpha + b\beta)\frac{\partial \tilde{u}}{\partial \tilde{x}} = 0.$$

The idea is to follow the subcharacteristics; therefore we choose $\tilde{x}(x, t)$ as a solution of the initial value problem:

$$a(x, t)\frac{\partial \tilde{x}}{\partial t} + b(x, t)\frac{\partial \tilde{x}}{\partial x} = 0, \qquad \tilde{x}(x, 0) = x. \tag{2.7}$$

Let

$$B(\tilde{x}, \tau, \varepsilon) = -(\alpha + c_1\beta)(\alpha + c_2\beta)(x, t),$$
$$C(\tilde{x}, \tau, \varepsilon) = L_2[\tilde{x}](x, t),$$
$$D(\tilde{x}, \tau, \varepsilon) = d_1(x, t, \varepsilon), \qquad F(\tilde{x}, \tau, \varepsilon) = f_1(x, t, \varepsilon),$$
$$\tilde{a}(\tilde{x}, \tau, \varepsilon) = a(x, t), \qquad E(\tilde{x}, \tau, \varepsilon) = (2\alpha + (c_1 + c_2)\beta)(x, t), \tag{2.8}$$

and let $B_{\tilde{x}\tilde{x}}$ be the operator

$$B_{\tilde{x}\tilde{x}}[u] = B\frac{\partial^2 u}{\partial \tilde{x}^2} - C\frac{\partial u}{\partial \tilde{x}} - Du.$$

We rewrite (2.5), in a simpler form:

$$-B_{\tilde{x}\tilde{x}}[\tilde{u}] + \tilde{a}\frac{\partial \tilde{u}}{\partial \tau} + \varepsilon E\frac{\partial^2 \tilde{u}}{\partial \tilde{x}\,\partial \tau} + \varepsilon^2 \frac{\partial^2 \tilde{u}}{\partial \tau^2} = F. \tag{2.9}$$

Moreover, we note with (2.7) that

$$B(\tilde{x}, \tau, \varepsilon) = -\left(-\frac{b}{a}+c_1\right)\left(-\frac{b}{a}+c_2\right)\beta^2.$$

Then, from the "timelike" hypothesis (2.3), we have

$$B(\tilde{x}, \tau, \varepsilon) > p^2\beta^2 \geqq 0, \tag{2.10}$$

and from the regularity of a, b and hypothesis (2.2), we deduce (see Appendix A.5) that $\beta \neq 0$. Now, we assume that

$$B = O_s(1), \qquad C = O(1). \tag{H}$$

Consequently, $B \geqq b_0 > 0$, b_0 constant. Then we have formally

$$-B_{\tilde{x}\tilde{x}}[\tilde{u}] + \tilde{a}\frac{\partial \tilde{u}}{\partial \tau} = F + O(\varepsilon). \tag{2.11}$$

And the first term of an asymptotic development of u for a large timescale is a solution of a nondegenerate parabolic problem:

$$-B_{\tilde{x}\tilde{x}}[\tilde{u}_0] + \tilde{a}\frac{\partial \tilde{u}_0}{\partial \tau} = F, \qquad \tilde{u}_0(\tilde{x}, 0) = g(\tilde{x}).$$

It is easy to see that (H) is equivalent to

$$\frac{\partial \tilde{x}}{\partial x} = O_s(1) \quad \text{and} \quad \frac{\partial^2 \tilde{x}}{\partial x^2} = O(1) \quad \text{uniformly,} \quad \text{for } x \in R, \quad t \in [0, T/\varepsilon];$$

$T > 0$ are arbitrary constant.

Thus our asymptotic analysis supposes that the subcharacteristics of P_ε do not converge or diverge or oscillate for large timescales (see Fig. 2 and the Appendix). The detailed calculations for translating (H) in terms of hypotheses on a and b are given in the Appendix. The cases of divergence or convergence need an entirely different analysis. This local analysis will be the purpose of another paper.

On the basis of hypothesis (H), we look for a generalised development of u in the form

$$\tilde{u}^n_{as} = \tilde{u}_0(\tilde{x}, \tau, \varepsilon) + \varepsilon\tilde{u}_1(\tilde{x}, \tau, \varepsilon) + \varepsilon^2\tilde{u}_2(\tilde{x}, \tau, \varepsilon) + \cdots.$$

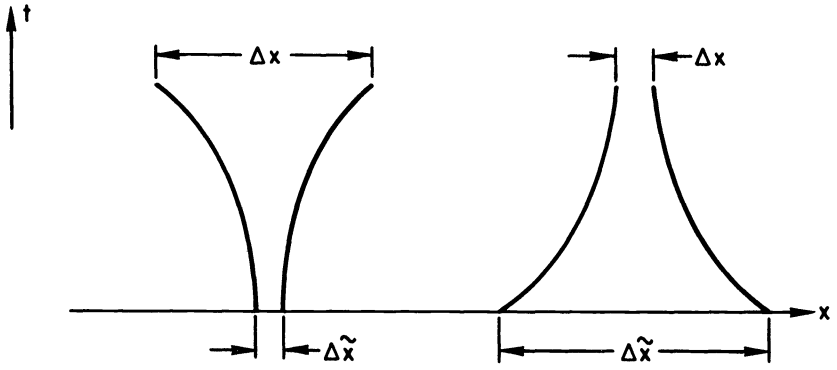


FIG. 2

Putting $\tilde{u}_{as}^n$, in (2.9), we find by identification in ε, the set of partial differential equations

$$
\begin{aligned}
&-B_{\tilde{x}\tilde{x}}[\tilde{u}_0]+\tilde{a}\frac{\partial \tilde{u}_0}{\partial \tau}=F,\\
&-B_{\tilde{x}\tilde{x}}[\tilde{u}_1]+\tilde{a}\frac{\partial \tilde{u}_1}{\partial \tau}=-E\frac{\partial^2 \tilde{u}_0}{\partial \tilde{x}\,\partial \tau},
\end{aligned}
\tag{2.12}
$$

and more generally for $k>1$

$$
-B_{\tilde{x}\tilde{x}}[\tilde{u}_k]+\tilde{a}\frac{\partial \tilde{u}_k}{\partial \tau}=-E\frac{\partial^2 \tilde{u}_{k-1}}{\partial \tilde{x}\,\partial \tau}-\frac{\partial^2 \tilde{u}_{k-2}}{\partial \tau^2}.
$$

We complete $\tilde{u}_{as}^n$ by the initial correction layer (see [3, Chap. 3, pp. 67-68])

$$
u_{as}^n=\sum_{k=0}^{n}\varepsilon^k\tilde{u}_k+\sum_{k=0}^{n}\varepsilon^{k+1}\nu_k(x,\hat{t})
$$

where $\hat{t}=t/\varepsilon$. The ν_k are solutions of the ordinary differential equations

$$
\frac{\partial^2 \nu_k}{\partial \hat{t}^2}+a(x,0)\frac{\partial \nu_k}{\partial \hat{t}}=\Phi_k(x,\hat{t}),
\tag{2.13}
$$

where Φ_k depends on $\nu_0, \nu_1, \cdots, \nu_{k-1}$ and $\Phi_0\equiv 0$. It is easy to derive the following set of initial conditions:

$$
\begin{aligned}
&\tilde{u}_0(\tilde{x},0)=g(\tilde{x}),\\
&\frac{\partial}{\partial \hat{t}}\nu_0(x,0)=-\frac{b}{a}(x,0)\cdot\frac{\partial}{\partial \tilde{x}}\tilde{u}_0(\tilde{x},0)+h(x),\\
&\tilde{u}_k(\tilde{x},0)=-\nu_{k-1}(x,0), \qquad k\geqq 1,\\
&\frac{\partial}{\partial \hat{t}}\nu_k(x,0)=\frac{\partial}{\partial \tau}\tilde{u}_{k-1}(\tilde{x},0)-\frac{b}{a}(x,0)\frac{\partial}{\partial \tilde{x}}\tilde{u}_k(\tilde{x},0), \qquad k\geqq 1
\end{aligned}
\tag{2.14}
$$

and the classical condition for boundary layer functions: $\lim_{\hat{t}\to\infty}\nu_k=0$, $B_{\tilde{x}\tilde{x}}$ is a nondegenerate elliptic operator, and u_{as}^n is uniquely determined. We easily verify that the $\nu_k(x,\hat{t})$ are independent of ε and of the following form:

$$
\begin{aligned}
&\nu_k(x,\hat{t})=P_k(\hat{t},x)\exp(-a(x,0)\hat{t})\\
&\text{with } P_k \text{ a polynomial in } \hat{t} \text{ and coefficients depending on } x.
\end{aligned}
\tag{2.15}
$$

As a practical application, we remark that in the case where the characteristics of P_ε are "quasi-parallel" for large timescales, namely,

$$
\begin{aligned}
&a\left(x,\frac{t}{\varepsilon}\right)-a_0=o(1), \qquad b\left(x,\frac{\tau}{\varepsilon}\right)-b_0=o(1),\\
&c_i\left(x,\frac{\tau}{\varepsilon}\right)-c_{i,0}=o(1), \qquad a_0, b_0, c_{i,0} \text{ constants}
\end{aligned}
\tag{2.16}
$$

and if we suppose in addition that $d\equiv 0$, our generalised development $\tilde{u}_{as}^n$ becomes a classical development ($\tilde{u}_i$ independent of ε). More precisely, we have in the first approximation

$$
\tilde{x}\sim x-\frac{b_0}{a_0}t,
$$

or

$$\beta \sim 1, \qquad \alpha \sim -\frac{b_0}{a_0}.$$

We look for $\tilde{u}_{as}^n$ in the form

$$\tilde{u}_{as}^n = \sum_{k=0}^{n} \varepsilon^k \tilde{u}_k(\tilde{x}, \tau)$$

and instead of (2.12), suppose the $\tilde{u}_i$ are solutions of the partial differential equations

$$-\left(c_{1,0} - \frac{b_0}{a_0}\right)\left(c_{2,0} - \frac{b_0}{a_0}\right)\frac{\partial^2 \tilde{u}_k}{\partial \tilde{x}^2} + \frac{\partial \tilde{u}_k}{\partial \tau} = \psi_k \tag{2.17}$$

where ψ_k depends on $\tilde{u}_0, \tilde{u}_1, \cdots, \tilde{u}_{k-1}$. Therefore, we obtain explicitly $\tilde{u}_{as}$ in the case where the characteristics of P_ε are "quasi-parallel" and $d \equiv 0$ for a large timescale.

To conclude this formal analysis, we have, using hypothesis (H), uniformly for $x \in R$, $t \in [0, T/\varepsilon]$, T arbitrary constant:

$$L_\varepsilon[\tilde{u}_{as}^n(\tilde{x}, \tau, \varepsilon)] = O(\varepsilon^{n+2}) \quad \text{for } -\infty < \tilde{x} < \infty, \quad 0 \leqq \tau \leqq T \tag{2.18}$$

and

$$L_\varepsilon\left[\sum_{k=0}^{n} \varepsilon^{k+1} \nu_k(x, \hat{t})\right] = O(\varepsilon^{n+1}) \quad \text{for } -\infty < x < \infty, \quad 0 \leqq \hat{t} \leqq \frac{T}{\varepsilon^2}. \tag{2.19}$$

Then

$$L_\varepsilon[u_{as}^n] = O(\varepsilon^{n+1}). \tag{2.20}$$

Moreover,

$$u_{as}^n(x, 0, \varepsilon) = g(x) + \varepsilon^{n+1} \nu_n(x, 0), \tag{2.21}$$

$$\frac{\partial}{\partial t} u_{as}^n(x, 0, \varepsilon) = h(x) + \varepsilon^{n+1} \frac{\partial \tilde{u}_n}{\partial \tau}(\tilde{x}, 0, \varepsilon). \tag{2.22}$$

Consequently, u_{as}^n is a uniform formal approximation of u, for $x \in R$ and $t \in [0, T/\varepsilon]$, T arbitrary constant.

3. Proof of validity. To prove the validity of our formal approximation, we look for a priori estimates for the solution $u(\tilde{x}, \tau, \varepsilon)$ of the following initial value problem:

$$\begin{cases} \tilde{L}(u) = -B\dfrac{\partial^2 u}{\partial \tilde{x}^2} + C\dfrac{\partial u}{\partial \tilde{x}} + \tilde{a}\dfrac{\partial u}{\partial \tau} + Du + \varepsilon E\dfrac{\partial^2 u}{\partial \tilde{x}\,\partial \tau} + \varepsilon^2 \dfrac{\partial^2 u}{\partial \tau^2} = \hat{F}(\tilde{x}, \tau, \varepsilon), \\ u(\tilde{x}, 0, \varepsilon) = \hat{g}(\tilde{x}), \qquad \tilde{x} \in R, \\ \dfrac{\partial u}{\partial \tau}(\tilde{x}, 0, \varepsilon) = \hat{h}(\tilde{x}), \qquad \tilde{x} \in R. \end{cases} \tag{3.1}$$

B, C, D, E, $\tilde{a}$ are defined in (2.8) and

$$B \geqq b_0 > 0, \qquad \tilde{a} \geqq a_0 > 0. \tag{3.2}$$

Furthermore, B, C, $\tilde{a}$, D, E are smooth uniformly bounded functions for $x \in R$, $0 \leqq \tau \leqq T$.

The main body of this section is the proof of the following a priori estimates.

LEMMA. *For ε sufficiently small, the solution $u(\tilde{x}, \tau, \varepsilon)$ of the initial value problem* (3.1) *satisfies the following pointwise estimate (uniformly valid in Ω):*

$$\sup_{\Omega} |u(\tilde{x}, \tau, \varepsilon)| \leqq C(\Omega) \exp\left(\frac{M\tau}{m}\right) K(\Omega),$$

$$\sup_{\Omega} \left|\frac{\partial u}{\partial \tilde{x}}(\tilde{x}, \tau, \varepsilon)\right| \leqq \varepsilon^{-1/2} C_1(\Omega) K'(\Omega),$$

$$\sup_{\Omega} \left|\frac{\partial u}{\partial \tau}(\tilde{x}, \tau, \varepsilon)\right| \leqq \varepsilon^{-3/2} C_1(\Omega) K'(\Omega),$$

with $K(\Omega) = \|\hat{F}\|_{\Omega} + \|\hat{g}_{\tilde{x}}\|_I + \varepsilon \|\hat{h}\|_I + \|\hat{g}\|_I$, where $\|\ \|$ is the L^2 norm on the given domain, and with $K'(\Omega) = \sup_{\Omega}(\hat{F}) + \varepsilon^{1/2} \sup_I |\hat{g}_{\tilde{x}}| + \varepsilon^{3/2} \sup_I |\hat{h}| + K(\Omega)$, where Ω is a part of a strip of the upper plane bounded by two characteristics: $\hat{x} = \hat{\gamma}_1(t)$ and $\hat{x} = \hat{\gamma}_2(t)$ (see Fig. 4), where $I = [A_\varepsilon, B_\varepsilon]$.

To establish these results we will use the classical energy integral method and we multiply the differential equation (3.1) by the expression

$$\theta \cdot \frac{\partial u}{\partial \tau} + \delta \frac{\partial u}{\partial \tilde{x}} + \eta \cdot u \tag{3.3}$$

where the functions θ, δ, η must be determined; the result is the following expression:

$$\begin{aligned}
&\frac{1}{2}\frac{\partial}{\partial \tilde{x}}\{-B\delta u_{\tilde{x}}^2 - \varepsilon(E\theta - \varepsilon\delta)u_\tau^2 + (C\eta + D\delta)u^2 \\
&\qquad -2B\theta u_{\tilde{x}} \cdot u_\tau - 2B\eta u_{\tilde{x}} \cdot u + 2\varepsilon E\eta u_\tau \cdot u\} \\
&+\frac{1}{2}\frac{\partial}{\partial \tau}\left\{(B\theta + \varepsilon E\delta)u_{\tilde{x}}^2 + \varepsilon^2 \theta u_\tau^2 + \left(\tilde{a}\eta + \theta D - 2\varepsilon \frac{\partial}{\partial \tilde{x}}(E\eta)\right)u^2 \right. \\
&\qquad \left. +2\varepsilon^2 \delta u_\tau u_{\tilde{x}} + 2\varepsilon^2 \eta u_\tau \cdot u\right\} \\
&+\left(\left\{\boxed{-\frac{1}{2}\frac{\partial}{\partial \tau}(B\theta)} + B\eta + \frac{\partial}{\partial \tilde{x}}\left(\frac{B\delta}{2}\right) + C\delta - \varepsilon \frac{\partial}{\partial \tau}\left(\frac{E \cdot \delta}{2}\right)\right\} u_{\tilde{x}}^2\right. \\
&\quad +\left\{-\varepsilon^2 \eta + \tilde{a}\theta - \varepsilon \frac{\partial}{\partial \tilde{x}}\left(\frac{E\theta}{2}\right) - \frac{1}{2}\varepsilon^2 \frac{\partial \theta}{\partial \tau} + \frac{\varepsilon^2}{2}\frac{\partial \delta}{\partial \tilde{x}}\right\} u_\tau^2 \\
&\quad +\left\{\boxed{-\frac{1}{2}\frac{\partial}{\partial \tau}(\theta D + \tilde{a}\eta)} + \frac{1}{2}\varepsilon \frac{\partial}{\partial \tau}\left(\frac{\partial}{\partial \tilde{x}}(E\eta)\right) + d\eta - \frac{\partial}{\partial \tilde{x}}\left(\frac{CD}{2}\right) - \frac{\partial}{\partial \tilde{x}}\left(\frac{D\delta}{2}\right)\right\} u^2 \\
&\quad \left. +\left\{\boxed{\frac{\partial}{\partial \tilde{x}}(B\theta) + C\theta + \tilde{a}\delta} - \varepsilon E\eta - \varepsilon^2 \frac{\partial \delta}{\partial \tau}\right\} u_{\tilde{x}} \cdot u_\tau + \left\{-\varepsilon^2 \frac{\partial}{\partial \tau}(\eta)\right\} u \cdot u_\tau\right) \\
&= \hat{F} \cdot (\delta u_{\tilde{x}} + \theta u_\tau + \eta u).
\end{aligned} \tag{3.4}$$

The framed terms are some leading terms a priori of order ε^{-1} that are important in the following.

We write (3.4) in the form

$$\frac{1}{2}\frac{\partial}{\partial \tau} Q_1 + \frac{1}{2}\frac{\partial}{\partial \tilde{x}} Q_2 - Q_3 = \hat{F} \cdot (\delta u_{\tilde{x}} + \theta u_\tau + \eta u). \tag{3.5}$$

In a first approach, we could assume

$$\hat{F} \in L^{\infty}(0, T, L^2(R)), \quad \hat{g} \in H^1(R), \quad \hat{h} \in L^2(R)$$

and integrate (3.5) in $R \times [0, T]$. Under these conditions, we can derive an estimate of the following form (using the $L^{\infty}(0, T, L^2(R))$ norm):

$$\left|\frac{\partial u}{\partial \tilde{x}}\right| + \varepsilon \left|\frac{\partial u}{\partial \tau}\right| + |u| \leqq C \exp(k\tau)|F| \tag{3.6}$$

and complete (3.6) in a trivial way using the initial conditions. In what follows we use a more complicated technique that also gives estimates in the maximum norm for $\partial u/\partial \tilde{x}$ and $\partial u/\partial \tau$. Moreover, we only need to assume that

$$\hat{F} \in C^0(R \times [0, T]), \quad \hat{g} \in C^1(R), \quad \hat{h} \in C^0(R)$$

and

$$\hat{F}, \hat{g}, \hat{g}_{\tilde{x}}, \hat{h} \quad \text{uniformly bounded} \quad \text{for } \tilde{x} \in R, \quad 0 \leqq \tau \leqq T.$$

The idea is to work in the domain of dependence. More precisely, we rewrite (3.1) in the new form:

$$\varepsilon^2 \left\{ \frac{\partial^2 u}{\partial \tau^2} + (c_1^* + c_2^*) \frac{\partial^2 u}{\partial \tilde{x}\, \partial \tau} + c_1^* \cdot c_2^* \frac{\partial^2 u}{\partial \tilde{x}^2} \right\} + \tilde{a} \frac{\partial u}{\partial \tau} + C \cdot \frac{\partial u}{\partial \tilde{x}} + D \cdot u = \hat{F}(\tilde{x}, \tau, \varepsilon) \tag{3.7}$$

where

$$c_1^* = (\alpha + c_1\beta)/\varepsilon \quad \text{and} \quad c_2^* = (\alpha + c_2\beta)/\varepsilon.$$

It is easy to see that the characteristics of the second-order differential operator in (3.7) are the same as the characteristics of L_2. For ε sufficiently small, "timelike behavior" is verified, namely, with (2.3) and (2.7),

$$\hat{p} + c_1^* < \frac{C}{\tilde{a}} < c_2^* - \hat{p}, \qquad \hat{p} > 0, \quad \hat{p} \text{ constant, independent of } \varepsilon. \tag{3.8}$$

We remark immediately that for a given $M(\tilde{x}, \tau)$, $\tau > 0$, the extent of the domain of dependence of the initial conditions is of order $O(\varepsilon^{-1})$ (see Fig. 3).

Now, we can apply the technique used in Geel [3, p. 53] and integrate (3.5) in Ω (instead of $R \times [0, T]$) where Ω is a part of a strip of the upper plane bounded with two characteristics $\tilde{x} = \gamma_1(t)$ and $\tilde{x} = \gamma_2(t)$ (see Fig. 4).

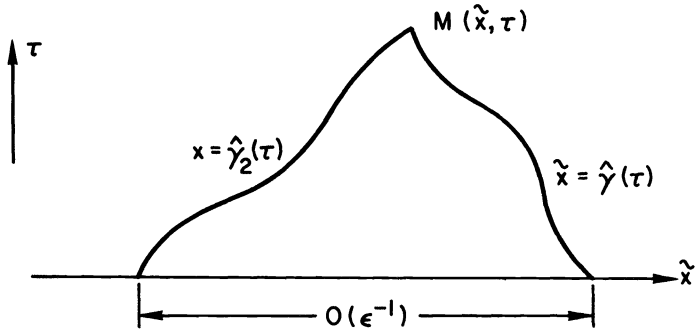


FIG. 3

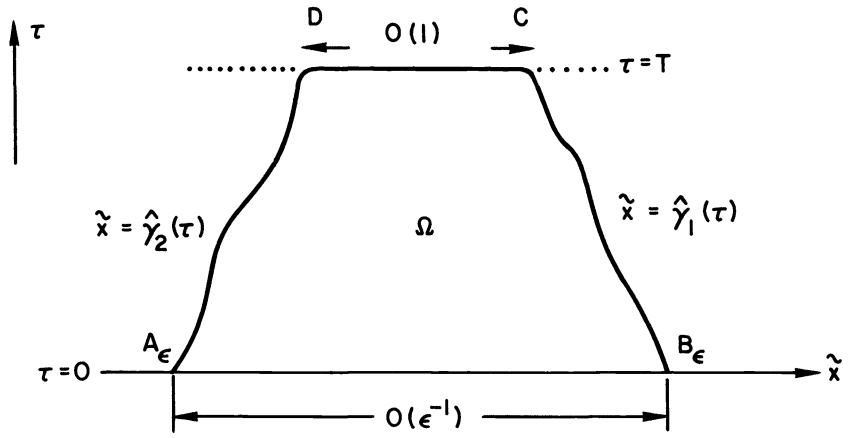


FIG. 4

Using Green's theorem, we obtain

$$\frac{1}{2}\left[\int_D^C Q_1\,d\tilde{x}+\int_{B_\varepsilon}^C (Q_2-c_1^*Q_1)\,d\tau+\int_{A_\varepsilon}^D (c_2^*Q_1-Q_2)\,d\tau\right]$$
$$=\iint_\Omega Q_3\,d\tilde{x}\,d\tau+\iint_\Omega \hat{F}(\delta\cdot u_{\tilde{x}}+\theta\cdot u_\tau+\eta\cdot u)\,d\tilde{x}\,d\tau+\int_{A_\varepsilon}^{B_\varepsilon} Q_1\,d\tilde{x}. \tag{3.9}$$

Then, for ε sufficiently small, we must choose the functions δ, θ, η such that Q_1, $Q_2-c_1^*Q_1$, and $c_2^*Q_1-Q_2$ are quadratic in u, u_τ, and $u_{\tilde{x}}$ and positive definite in Ω, and we have good majorants for $|Q_1|$ and $|Q_3|$, by quadratic positive-definite expressions in u, u_τ, $u_{\tilde{x}}$ in Ω. To reach this end, we put (see (3.4))

$$B\cdot\theta=m_1, \qquad m_1 \text{ constant},$$
$$\theta\cdot D+\tilde{a}\eta=m_2, \qquad m_2 \text{ constant}$$
$$\partial/\partial\tilde{x}(B\cdot\theta)+C\theta+\tilde{a}\delta=0,$$

so we have

$$\theta=m_1\cdot B^{-1},$$
$$\eta=(m_2-m_1\cdot B^{-1}\cdot D)\tilde{a}^{-1}, \tag{3.10}$$
$$\delta=-(C\cdot B^{-1}\cdot m_1)\tilde{a}^{-1},$$

and the constants m_1, m_2 are left free.

We now have q_0 and M positive and independent of ε such that for ε sufficiently small

$$-Q_3=\left(\tilde{a}\theta-\varepsilon\frac{\partial}{\partial\tilde{x}}\left(\frac{E\theta}{2}\right)\right)u_\tau^2-\hat{Q}_3 \tag{3.11}$$

and

$$|-\hat{Q}_3|\leqq M(u_{\tilde{x}}^2+\varepsilon^2u_\tau^2+u^2) \quad \text{uniformly in } \Omega \tag{3.12}$$

and

$$\tilde{a}\theta - \varepsilon \frac{\partial}{\partial \tilde{x}}\left(\frac{E\theta}{2}\right) > q_0 > 0. \tag{3.13}$$

We also have (with $m_1, m_2 > 0$), M_1, m positive, independent of ε, such that for ε sufficiently small:

$$Q_1 > m(u_{\tilde{x}}^2 + \varepsilon^2 u_\tau^2 + u^2) \tag{3.14}$$

$$Q_1 < M_1(u_{\tilde{x}}^2 + \varepsilon^2 u_\tau^2 + u^2) \qquad \text{uniformly in } \Omega. \tag{3.15}$$

Finally, we easily verify that we can choose m_1, m_2 positive constants such that there exists a positive constant m' independent of ε with

$$\begin{aligned} &Q_2 - c_1^* Q_1 > \frac{m'}{\varepsilon}(u_{\tilde{x}}^2 + \varepsilon^2 u_\tau^2 + u^2) \\ &c_2^* Q_1 - Q_2 > \frac{m'}{\varepsilon}(u_{\tilde{x}}^2 + \varepsilon^2 u_\tau^2 + u^2) \end{aligned} \qquad \text{uniformly in } \Omega. \tag{3.16}$$

From (3.9) we now deduce the estimate

$$\begin{aligned} &m\int_D^C (u_{\tilde{x}}^2 + \varepsilon^2 u_\tau^2 + u^2)\, d\tilde{x} + \frac{m'}{\varepsilon}\int_{B_\varepsilon}^C (u_{\tilde{x}}^2 + \varepsilon^2 u_\tau^2 + u^2)\, d\tau \\ &\quad + \frac{m}{\varepsilon}\int_{A_\varepsilon}^D (u_{\tilde{x}}^2 + \varepsilon^2 u_\tau + u^2)\, d\tau + q_0 \iint_\Omega (u_{\tilde{x}}^2 + u_\tau^2 + u^2)\, d\tilde{x}\, d\tau \\ &\leqq M_2\left[\iint_\Omega (u_{\tilde{x}}^2 + \varepsilon^2 u_\tau^2 + u^2)\, d\tilde{x}\, d\tau + \left(\iint_\Omega F^2\, d\tilde{x}\, d\tau\right)^{1/2}\right. \\ &\quad \left. \cdot \left(\iint_\Omega (\tilde{u}_x^2 + u_\tau^2 + u^2)\, d\tilde{x}\, d\tau\right)^{1/2} + \int_{A_\varepsilon}^{B_\varepsilon} (u_{\tilde{x}}^2 + \varepsilon^2 u_\tau^2 + u^2)\, d\tilde{x}\right]. \end{aligned} \tag{3.17}$$

Note that this formula is also valid for any similar region $\Omega' \subset \Omega$ (with the same constant) (see Fig. 5).

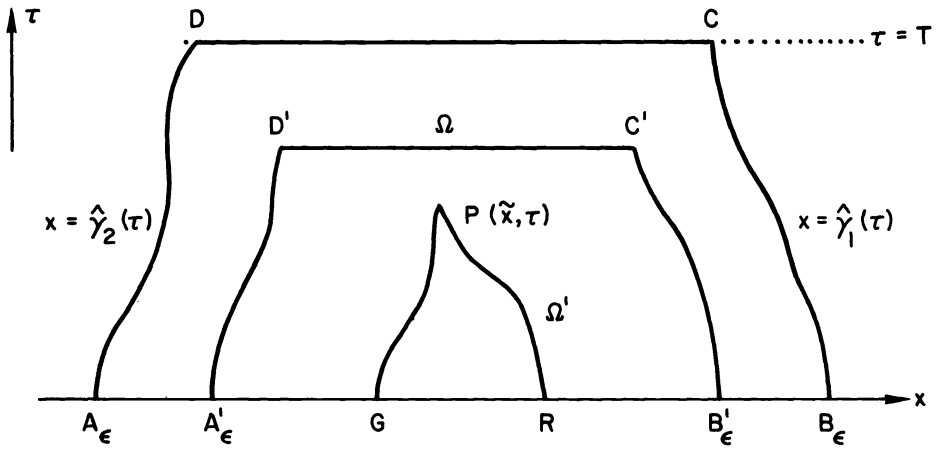


FIG. 5

Next we deduce

$$\int_{\hat{\gamma}_2(\tau)}^{\hat{\gamma}_1(\tau)} (u_{\tilde{x}}^2+\varepsilon^2 u_\tau^2+u^2)\,d\tilde{x}-\frac{M}{m}\int_0^\tau \int_{\hat{\gamma}_2(\tau)}^{\hat{\gamma}_1(\tau)} (u_{\tilde{x}}^2+\varepsilon^2 u_\tau^2+u^2)\,d\tilde{x}\,d\tau$$

$$\leqq \frac{M}{m}\left(\iint_\Omega F^2\,d\tilde{x}\,d\tau\right)^{1/2}\left(\iint_\Omega (u_{\tilde{x}}^2+u_\tau^2+u^2)\,d\tilde{x}\,d\tau\right)^{1/2} \tag{3.18}$$

$$+\int_{\Lambda_\varepsilon}^{B_\varepsilon} (u_{\tilde{x}}^2+\varepsilon^2 u_\tau^2+u^2)\,d\tilde{x}.$$

Let

$$K(\Omega)=\|F\|_\Omega+\|\hat{g}_{\tilde{x}}\|_I+\varepsilon\|\hat{k}\|_I+\|\hat{g}\|_I.$$

Then using Gronwall's lemma, we derive from (3.18) the estimate

$$\int_{\hat{\gamma}_2(t)}^{\hat{\gamma}_1(t)} (u_{\tilde{x}}^2+\varepsilon^2 u_\tau^2+u^2)\,d\tilde{x}\leqq\frac{M'}{m}\exp\left(\frac{M\tau}{m}\right)K^2(\Omega) \tag{3.19}$$

and it is easy to obtain the pointwise estimate, uniformly in Ω

$$|u(\tilde{x},\tau,\varepsilon)|<C(\Omega)\exp\left(\frac{M\tau}{m}\right)K(\Omega) \tag{3.20}$$

where $C(\Omega)$ is independent of ε. Applying (3.17) to a region Ω' instead of Ω (see Fig. 5) and using the estimate (3.19), we have

$$\frac{1}{\varepsilon}\int_{B'_\varepsilon}^{C'} (u_{\tilde{x}}^2+u_\tau^2+u^2)\,d\tau\leqq\text{const.}\exp\left(\frac{M\tau}{m}\right)K^2(\Omega), \tag{3.21}$$

$$\frac{1}{\varepsilon}\int_{\Lambda'_\varepsilon}^{D'} (u_{\tilde{x}}^2+\varepsilon^2 u_\tau^2+u^2)\,d\tau\leqq\text{const.}\exp\left(\frac{M\tau}{m}\right)K^2(\Omega). \tag{3.22}$$

We finally deduce a pointwise estimate for the first derivative of u.

Using (3.7), we have easily

$$\varepsilon^2\left(\frac{\partial}{\partial\tau}+c_2^*\frac{\partial}{\partial\tilde{x}}\right)[(u_\tau+c_1^*u_{\tilde{x}})^2]\leqq F^2+\text{const.}\left(u_\tau^2+\frac{1}{\varepsilon^2}u_{\tilde{x}}^2+u^2\right). \tag{3.23}$$

Integrating this inequality with respect to τ, along the characteristic QP and using (3.22) (see Fig. 4), we obtain

$$|(u_\tau+c_1^*u_{\tilde{x}})(x,\tau,\varepsilon)|<\max_I|\hat{h}|+\varepsilon^{-1}\max_I|g_{\tilde{x}}|+\varepsilon^{-1/2}\,\text{const.}\exp\left(\frac{M\tau}{m}\right)\cdot K(\Omega)$$

$$+\varepsilon^{-3/2}\max_\Omega|F|.$$

Or

$$|(u_\tau+c_1^*u_{\tilde{x}})(\tilde{x},\tau,\varepsilon)|<\varepsilon^{-3/2}C'(\Omega)K'(\Omega) \tag{3.24}$$

where $C'(\Omega)$ is independent of ε and

$$K'(\Omega)=\left(\varepsilon^{3/2}\sup_I|\hat{h}|+\varepsilon^{1/2}\sup_I|\hat{g}_{\tilde{x}}|+\text{const.}\,K(\Omega)+\sup_\Omega|\hat{F}|\right).$$

By an entirely analogous method we obtain

$$|(u_\tau + c_2^* u_{\tilde{x}})(\tilde{x}, \tau, \varepsilon)| \leqq \varepsilon^{-3/2} C'(\Omega) K'(\Omega) \quad \text{uniformly in } \Omega. \tag{3.25}$$

We have thus demonstrated all results stated in the lemma. □

To conclude this section, we apply our a priori estimate to the formal approximation $u_{as}^n(x, t, \varepsilon)$ constructed in the previous section.

Let $z(x, t, \varepsilon) = (u - u_{as}^n)(x, t, \varepsilon) = \tilde{z}(\tilde{x}, \tau, \varepsilon)$. We have with (2.18)-(2.22):

$$\begin{aligned} &|\tilde{L}_\varepsilon[\tilde{z}]| < \text{const. } \varepsilon^{n+1} \quad \text{uniformly in } R \times [0, T], \\ &|\tilde{z}(\tilde{x}, 0)| < \text{const. } \varepsilon^{n+1} \quad \text{uniformly in } R, \\ &\left|\frac{\partial \tilde{z}}{\partial \tau}(\tilde{x}, 0)\right| < \text{const. } \varepsilon^{n} \quad \text{uniformly in } R, \\ &\left|\frac{\partial \tilde{z}}{\partial \tilde{x}}(\tilde{x}, 0)\right| < \text{const. } \varepsilon^{n+1} \quad \text{uniformly in } R. \end{aligned} \tag{3.26}$$

Then

$$K(\Omega) \leqq \text{const. } \varepsilon^{n+1/2}, \qquad K'(\Omega) \leqq \text{const. } \varepsilon^{n+1/2}$$

and

$$\begin{aligned} &\sup_\Omega |z(\tilde{x}, t) = O(\varepsilon^{n+1/2}), \\ &\sup_\Omega \left|\frac{\partial z}{\partial \tau}(\tilde{x}, t)\right| = O(\varepsilon^n), \\ &\sup_\Omega \left|\frac{\partial z}{\partial \tilde{x}}(\tilde{x}, t)\right| = O(\varepsilon^n). \end{aligned}$$

Using the fact that, for arbitrary $m > n$:

$$\begin{aligned} &\sup_\Omega |u_{as}^m - u_{as}^n| = O(\varepsilon^{n+1}), \\ &\sup_\Omega \left|\frac{\partial}{\partial x} u_{as}^m - \frac{\partial}{\partial x} u_{as}^n\right| = O(\varepsilon^{n+1}), \\ &\sup_\Omega \left|\frac{\partial}{\partial t} u_{as}^m - \frac{\partial}{\partial t} u_{as}^n\right| = O(\varepsilon^{n+1}), \end{aligned}$$

we derive the following theorem.

THEOREM. *Let* $D_\varepsilon = \{(x, t), (\tilde{x}, \tau) \in \Omega\}$. *Under the previous hypothesis and definitions, we have the following approximation, uniformly valid in* D_ε *for arbitrary integers* $n > 0$:

$$\begin{aligned} u_\varepsilon(x, t) &= \sum_{k=0}^{n} \varepsilon^k \tilde{u}_k(\tilde{x}, \tau) + \sum_{k=0}^{n-1} \varepsilon^{k+1} \nu_k(x, \hat{t}) + O(\varepsilon^{n+1}), \\ \frac{\partial}{\partial x} u_\varepsilon(x, t) &= \beta \sum_{k=0}^{n} \varepsilon^k \frac{\partial \tilde{u}_k}{\partial \tilde{x}}(\tilde{x}, \tau) + \sum_{k=0}^{n-1} \varepsilon^{k+1} \frac{\partial}{\partial x} \nu_k(x, \hat{t}) + O(\varepsilon^{n+1}), \\ \frac{\partial}{\partial t} u_\varepsilon(x, t) &= \alpha \sum_{k=0}^{n} \varepsilon^k \frac{\partial}{\partial \tilde{x}} \tilde{u}_k(\tilde{x}, \tau) + \sum_{k=0}^{n-1} \varepsilon^{k+1} \frac{\partial}{\partial \tau} \tilde{u}_k(\tilde{x}, \tau) \\ &\quad + \sum_{k=0}^{n} \varepsilon^k \frac{\partial}{\partial \hat{t}} \nu_k(x, \hat{t}) + O(\varepsilon^{n+1}). \end{aligned}$$

We also have (outside the initial boundary layer) in $D_\varepsilon^* = D_\varepsilon \setminus (R \times [0, t_0])$, *where* t_0 *is an arbitrary small positive constant independent of* ε, *the same estimate without the boundary correction layer. In particular, we have*

$$u_\varepsilon(x, t) = \tilde{u}_0(\tilde{x}, \tau) + O(\varepsilon),$$

$$\frac{\partial}{\partial x} u_\varepsilon(x, t) = \beta \frac{\partial}{\partial \tilde{x}} \tilde{u}_0(\tilde{x}, \tau) + O(\varepsilon) \quad \textit{uniformly valid in } D_\varepsilon^*,$$

$$\frac{\partial}{\partial t} u_\varepsilon(x, t) = \alpha \frac{\partial}{\partial \tilde{x}} \tilde{u}_0(\tilde{x}, \tau) + O(\varepsilon).$$

Remark 1. We note that the estimate in the maximum norm for u of the lemma is optimal: To show this for simplicity we take constant coefficients in P_ε. Let us suppose in addition that

$$\varepsilon \frac{\partial^2 \tilde{u}_0}{\partial \tau \, \partial \tilde{x}} + \varepsilon \frac{2\partial^2 \tilde{u}_0}{\partial \tau^2} \in L^2(0, T; L^2(R)).$$

Then for $z = u - \tilde{u}_0$, we have $K(\Omega) = O(\varepsilon)$ and the lemma implies that

$$u - \tilde{u}_0 = O(\varepsilon).$$

Remark 2. Our result could be generalised to quasilinear problems of the following form:

$$\varepsilon \left(\frac{\partial}{\partial t} + c_1 \frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial t} + c_2 \frac{\partial}{\partial x}\right)[u] + a(x, t, \varepsilon) \frac{\partial u}{\partial t} + b(x, t, \varepsilon) \frac{\partial u}{\partial x} + d(x, t, u, \varepsilon) = 0$$

with d of order ε for large time.

Formally, we have instead of (2.11), with $d(x, t, u, \varepsilon) = \varepsilon \tilde{d}_1(\tilde{x}, \tau, \tilde{u}, \varepsilon)$,

$$-B_{\tilde{x}\tilde{x}}[\tilde{u}] + \tilde{a} \frac{\partial \tilde{u}}{\partial \tau} + \tilde{d}_1(\tilde{x}, \tau, \tilde{u}, \varepsilon) = O(\varepsilon),$$

and following Hsiao and Weinacht [7], we must suppose that the reduced problem has a bounded classical solution (for $\tau \leqq T$).

To prove the validity, we can use a fixed point theorem [3, Chap. 2], with our a priori estimate for the linearized operator:

$$\varepsilon \left(\frac{\partial}{\partial t} + c_1 \frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial t} + c_2 \frac{\partial}{\partial x}\right)[u] + a(x, t, \varepsilon) \frac{\partial u}{\partial t} + b(x, t, \varepsilon) \frac{\partial u}{\partial x} + d(x, t, 0, \varepsilon) \cdot u.$$

Appendix. In this part we present some calculations for translating (H) into conditions on the coefficients a and b in L_1. First we will derive explicit formulas for α and β.

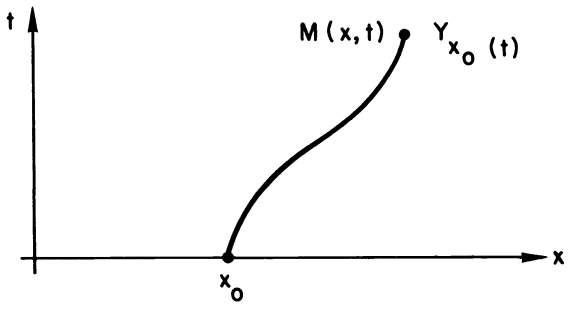


FIG. 6

Let $Y(x_0, t)$ be a characteristic of L_1 (see Fig. 6):

$$\text{(A1)} \qquad \begin{aligned} &\frac{\partial}{\partial t} Y(x_0, t) = \frac{b}{a}(Y(x_0, t), t), \\ &Y(x_0, 0) = x_0. \end{aligned}$$

We know that $\tilde{x}(Y(x_0, t), t) = x_0$, and

$$\text{(A2)} \qquad \alpha = \frac{\partial \tilde{x}}{\partial t} = -\frac{b}{a}\frac{\partial \tilde{x}}{\partial x} = -\frac{b}{a}\beta \quad \text{with } a \geqq a_0 > 0.$$

It is sufficient to obtain β; we have that

$$\text{(A3)} \qquad \beta = \frac{\partial \tilde{x}}{\partial x} = \frac{\partial x_0}{\partial x} \quad \text{and} \quad \frac{\partial x}{\partial x_0} = \frac{\partial Y(x_0, t)}{\partial x_0}$$

is the solution of

$$\text{(A4)} \qquad \begin{aligned} &\frac{\partial}{\partial t}\left(\frac{\partial Y(x_0, t)}{\partial x_0}\right) = \left(\frac{a}{b}\right)_1 (Y(x_0, t), t) \cdot \frac{\partial Y(x_0, t)}{\partial x_0}, \\ &\frac{\partial Y(x_0, 0)}{\partial x_0} = 1 \end{aligned}$$

where we note $(b/a)_1(x, t) = \partial/\partial x[(b/a)(x, t)]$. Then, we have the formula

$$\text{(A5)} \qquad \beta = \frac{\partial \tilde{x}}{\partial x} = \exp -\left\{\int_0^t \left(\frac{b}{a}\right)_1 (Y(\tilde{x}, \xi), \xi) \cdot d\xi\right\}.$$

(Note that $\beta \neq 0$!.) Let

$$A_0^t = \int_0^t \left(\frac{b}{a}\right)_1 (Y(\tilde{x}, \xi), \xi)\, d\xi$$

and

$$B_0^t = \int_0^t \left(\frac{b}{a}\right)_{11} (Y(\tilde{x}, \xi), \xi)\, d\xi$$

where we note that $(b/a)_{11}(x, t) = \partial^2/\partial x^2[(b/a)(x, t)]$. We finally have (H) equivalent to

$$\text{(A6)} \qquad \begin{aligned} &A_0^t = O(1) \\ &B_0^t = O(1) \end{aligned} \quad \text{uniformly for } x \in R,\ t \in [0, T/\varepsilon].$$

As an illustration, let us choose the simple situation: $a \equiv 1$, $b \equiv b(x)$. On a nonvertical characteristic, we have

$$\begin{aligned} &\frac{\partial \tilde{x}}{\partial x} \underset{t \to +\infty}{\sim} \exp \int^t b'(Y(\tilde{x}, \xi)) \cdot d\xi, \\ &\frac{\partial \tilde{x}}{\partial x} \underset{t \to +\infty}{\sim} \exp \int^t d(b(Y(\tilde{x}, \xi)))/b(Y(\tilde{x}, \xi)). \end{aligned}$$

Then

$$\text{(A7)} \qquad \frac{\partial \tilde{x}}{\partial x} \sim \text{const. } b^{-1}(Y(\tilde{x}, t)) \quad \text{and} \quad \frac{\partial^2 \tilde{x}}{\partial x^2} \sim \text{const. } b'(Y(\tilde{x}, t))/b^2(Y(\tilde{x}, t)).$$

If b is zero in x_0, on the vertical characteristic $x = x_0$, we have

$$\text{(A8)} \qquad \frac{\partial \tilde{x}}{\partial x} \underset{t \to +\infty}{\sim} \exp -b'(x_0) \cdot t$$

and if $b'(x_0)=0$: $\partial^2\tilde{x}/\partial x^2 \sim_{t\to+\infty} -b''(x_0)\cdot t$. Therefore, in the particular situation: $a\equiv 1$, $b\equiv b(x)$, (H) is verified if

$$|b| \geqq p > 0, \quad p \text{ constant.} \tag{A9}$$

If zero belongs to the closure of the range of b there are many special cases where (H) fails (see, for example, Fig. 7). We leave for a future publication the study of these special situations and their local asymptotic analysis.

More generally, (H) is satisfied if we assume that

$$\max_{x\in R}\left|\frac{\partial b\cdot a^{-1}}{\partial x}\right| + \max_{x\in R}\left|\frac{\partial^2 b\cdot a^{-1}}{\partial x^2}\right| \in L^1(R), \tag{A10}$$

but we can define a less restrictive hypothesis on the given functions a and b, when b/a is decomposed into

$$\frac{b}{a}(x,t) = b_0(x)\cdot l(t). \tag{A11}$$

On a nonvertical characteristic

$$\frac{\partial\tilde{x}}{\partial x} \sim \text{const. } b_0^{-1}(Y(\tilde{x},t)) \quad \text{and} \quad \frac{\partial^2\tilde{x}}{\partial x^2} \sim \text{const. } b_0'\cdot b_0^{-2}(Y(\tilde{x},t)) \tag{A12}$$

if b is zero in $x=x_0$, on the vertical characteristic $x=x_0$, we have

$$\frac{\partial\tilde{x}}{\partial x} \underset{t\to+\infty}{\sim} \exp -\int^t b_0'(x_0)l(\xi)\,d\xi, \tag{A13}$$

and if $b'(x_0)=0$: $\partial^2\tilde{x}/\partial x^2 \sim \int^t b_0''(x_0)l(\xi)\,d\xi$. Therefore, (H) is satisfied if

$$|b_0| \geqq p > 0 \quad \text{or} \quad l\in L^1(R). \tag{A14}$$

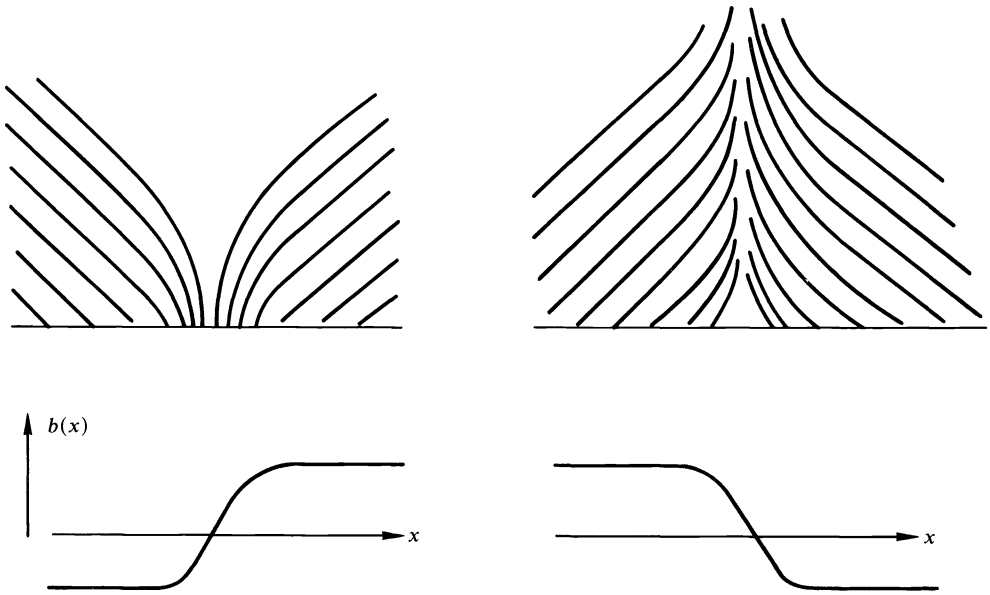


FIG. 7. *Field of characteristics of* L_1.

REFERENCES

[1] W. ECKHAUS, *Asymptotic Analysis of Singular Perturbations*, North-Holland, Amsterdam, 1979.

[2] ———, *The long-time behaviour for perturbed wave—equations and related problems*, preprint 404, Department of Mathematics, University of Utrecht, Utrecht, the Netherlands, 1985; published in part in Lecture Notes in Physics 246, Springer-Verlag, Berlin, New York, 1986.

[3] R. GEEL, *Singular perturbations of hyperbolic type*, thesis, University of Amsterdam, Amsterdam, the Netherlands, 1978.

[4] E. M. DE JAGER, *Singular perturbations of hyperbolic type*, Nieuw Arch. Wisk. 23 (1975), pp. 145-172.

[5] M. MADAUNE-TORT, *Perturbations singulières de problèmes aux limites du second ordre hyperboliques et paraboliques non linéaires*, Thèse, Université de Pau, Pau, France, 1982.

[6] G. B. WHITHAM, *Linear and Nonlinear Waves*, John Wiley, New York, 1974.

[7] G. C. HSIAO AND R. J. WEINACHT, *Singular perturbations for a semilinear hyperbolic equation*, SIAM J. Math. Anal., 14 (1983), pp. 1168-1179.

[8] A. BENAOUDA AND M. MADAUNE-TORT, *Hyperbolic-parabolic singular perturbation problems*, SIAM J. Math. Anal., 18 (1987), pp. 137-148.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 884-904, July 1990
© 1990 Society for Industrial and Applied Mathematics
004

GLOBAL STABILITY OF A PREMIXED REACTION ZONE (TIME-DEPENDENT LIÑAN'S PROBLEM)*

CARLOS ALVAREZ PEREIRA† AND JOSÉ M. VEGA†

Abstract. Global stability properties of a premixed, three-dimensional reaction zone are considered. In the nonadiabatic case (i.e., when there is a heat exchange between the reaction zone and the burned gases) there is a unique, spatially one-dimensional steady state that is shown to be unstable (respectively, asymptotically stable) if the reaction zone is cooled (respectively, heated) by the burned mixture. In the adiabatic case, there is a unique (up to spatial translations) steady state that is shown to be stable. In addition, the large-time asymptotic behavior of the solution is analyzed to obtain sufficient conditions on the initial data for stabilization. Previous partial numerical results on linear stability of one-dimensional reaction zones are thereby confirmed and extended.

Key words. global stability, stabilization, reaction regions, premixed flames, nonadiabatic flames

AMS(MOS) subject classifications. 35B35, 35B40, 35B55, 80A25, 80A32

1. Introduction. We consider the time-dependent structure of a premixed n-dimensional ($n=1, 2$, or 3) reaction zone which, after convenient nondimensionalization, is governed by

$$\partial u/\partial t = \Delta u - (u/2)\exp(mx_1 - u) \quad \text{for } (x, t) \in \mathbb{R}^n \times]0, T_0[, \tag{1.1}$$

$$u \to 0 \quad \text{if } m \leqq 0, \qquad u \text{ is bounded} \quad \text{if } m > 0 \quad \text{as } x_1 \to -\infty, \tag{1.2}$$

$$|u - x_1| \text{ is bounded} \quad \text{as } x_1 \to \infty, \tag{1.3}$$

$$u \text{ is bounded} \quad \text{as } x_2^2 + \cdots + x_n^2 \to \infty, \tag{1.4}$$

$$u(x, 0) = \varphi(x) > 0 \quad \text{for } x \in \mathbb{R}^n, \tag{1.5}$$

where conditions (1.2) and (1.3) are assumed to hold uniformly for $(x_2, \cdots, x_n) \in \mathbb{R}^{n-1}$ and for $t \in [0, T]$, for all $T \in [0, T_0[$ (for some $T_0 \leqq \infty$), and condition (1.4) is assumed to hold uniformly for $(x_1, t) \in I \times [0, T]$, for all bounded intervals $I \subset \mathbb{R}$ and all $T \in [0, T_0[$.

Here, Δ is the Laplacian operator, t and $x = (x_1, \cdots, x_n)$ ($n = 1, 2$, or 3) are the time and space variables, and $u \geqq 0$ is a reactant concentration. The parameter m is a measure of the heat flux (heat loss if $m > 0$ and heat gain if $m < 0$) from the reaction zone towards the burned mixture, which is located at $x_1 = -\infty$; m is assumed to satisfy $-\infty < m < 1$, for the chemical reaction to be frozen (i.e., for the reaction term $(u/2)\exp(mx_1 - u)$ to vanish) at the fresh mixture (i.e., at $x_1 = +\infty$). The initial state φ is assumed to satisfy the boundary conditions (1.2)-(1.4), which, of course, are expected to be superfluous; they are written to emphasize that the solution of the Cauchy problem (1.1)-(1.5) is physically meaningful only if it satisfies (1.2)-(1.4).

In this paper we will analyze the stability of steady states of (1.1)-(1.3) that depend only on the x_1 coordinate. Since the reaction term does not depend explicitly on the x_2 and x_3 coordinates, it makes sense (mathematically) to consider the (spatially) one- and two-dimensional cases in which $u = u(x_1, t)$ and $u = u(x_1, x_2, t)$, respectively. But

* Received by the editors April 3, 1989; accepted for publication (in revised form) August 24, 1989. This research was partially supported by the Comisión Interministerial de Ciencia y Tecnología under grant PB 86-0497.

† E.T.S. Ingenieros Aeronáuticos, Universidad Politécnica de Madrid, 28040 Madrid, Spain.

since the underlying physical problem is spatially three-dimensional, to obtain conclusive stability results we must consider (1.1)-(1.5) in three space dimensions. It is not at all obvious (although it will be true under certain conditions for (1.1)-(1.5)) that initial inhomogeneities in the x_2 and x_3 coordinates disappear as $t \to \infty$. Some results in the literature [1] could perhaps be extended to include (1.1)-(1.5) if the spatial domain $\mathbb{R}^3$ were replaced by a cylinder $\Omega = \mathbb{R} \times \Omega_1$, with $\Omega_1 \subset \mathbb{R}^2$ bounded, and if boundary conditions of the Neumann type were imposed on $\mathbb{R} \times \partial\Omega_1$, provided that the size of Ω_1 is sufficiently small. But to assume that the characteristic lengths in the x_2 and x_3 directions are small (or even finite) is not justified from a physical point of view. Therefore, we will consider (1.1)-(1.5) mainly for $n = 3$, although the case where $n = 1$ will be considered also for technical reasons.

The one-dimensional, time-independent version of (1.1)-(1.3) was introduced by Liñán [2], in a pioneering work on counterflow diffusion flames in the large activation energy limit, and (1.1)-(1.5) is currently known in the literature as Liñán's problem. It has subsequently appeared in high-activation energy analysis of many other realistic problems that are significant in both combustion and chemical reactor theory. For example, it has appeared in the analysis of burning monopropellant drops [3]-[5], chambered diffusion flames [6], two-step sequential reactions [7], [8], and tubular nonadiabatic chemical reactors [9]; in all these instances, the parameter m is different from zero, but the adiabatic case ($m = 0$) appears in a large number of problems [10], such as the analysis of premixed flames [11]-[13] and porous catalysts [14], [15], to cite only two examples.

Problem (1.1)-(1.5) is also of interest if the nonlinearity $u \exp(mx_1 - u)$ is replaced by a more general one. For example, if we use Langmuir-Hinshelwood kinetic laws for the chemical reaction, instead of the Arrhenius law that has been used to derive (1.1), we obtain nonlinearities of the type [16], [17]

$$u^p/(1+u)^q \quad \text{or} \quad [u^p/(a+u)^r] \exp(mx_1 - u), \tag{1.6}$$

where $p \geqq 0$, $q > p+1$, $a > 0$, and $m < 1$ (the exponents p, q, and r are not necessarily integers). These generalizations will be considered in remarks after some of the main results.

A numerical analysis of the one-dimensional steady states of (1.1)-(1.3) has been done by Liñán [2]. His results were rigorously proven true by Hastings and Poore [18], [19], who showed that the solution is unique if either $-\infty < m < 0$ or $0 < m < \frac{1}{2}$, while there is no solution if $\frac{1}{2} \leqq m < \infty$ (if $m = 0$, there is a unique steady state up to translations in the space variable, as is easily seen by means of simple phase-plane arguments). To derive stability results, we will need slightly more precise information about the dependence of the steady state on m for $0 < m < \frac{1}{2}$, which will be obtained in the Appendix, where a simpler proof of the results by Hastings and Poore [18], [19] (partially based on their ideas) will also be given.

The first analysis of the stability of the steady states of (1.1)-(1.4) is due to Peters [20], who computed numerically the maximum eigenvalue of the (self-adjoint) linearized problem in the spatially one-dimensional case, and found that $m > 0$ is necessary and sufficient for a strictly positive eigenvalue to exist. More recently, Stewart and Buckmaster [21] performed an asymptotic analysis of the same linearized problem in the limit $m \to 0^+$, which is singular. Those results ignore the continuous spectrum of the linearized problem, which has been calculated, for related spatially one-dimensional problems on combustion, by Buckmaster, Nachman, and Taliaferro [22], by means of a general theory developed by Taliaferro [23]. Unfortunately, Taliaferro's results deal with a weak notion of linear stability (a steady state is said to be stable if the maximum

of the spectrum is nonpositive and zero is not an eigenvalue) and, anyway, do not apply to the linearized problem associated with (1.1)-(1.3). Those results need completion also because they apply only to the one-dimensional case.

At this point, the boundary conditions (1.2), (1.3) deserve some attention. In this analysis of the steady state problem, Liñán [2] imposed the following conditions at $x_1 = \pm\infty$:

$$\partial u/\partial x_1 \to 0 \quad \text{as } x_1 \to -\infty, \qquad \partial u/\partial x_1 \to 1 \quad \text{as } x_1 \to \infty. \tag{1.7}$$

Stewart and Buckmaster [21] maintain conditions (1.7) for the time-dependent problem, while Peters [20] replaces them by

$$u \to c_1 \quad \text{as } x_1 \to -\infty, \qquad u - x_1 \to c_2 \quad \text{as } x_1 \to \infty \tag{1.8}$$

for some constants c_1 and c_2. In fact, conditions (1.2), (1.3), and (1.7) and (1.8), are equivalent when applied to the one-dimensional steady state problem (see the Appendix) and are seen to lead to equivalent linearized eigenvalue problems. But those three conditions are not equivalent when applied to the time-dependent problem. We will use conditions (1.2), (1.3), which are obtained from matching conditions in the singular perturbation analysis that leads to (1.1)-(1.5), as may be seen.

In this paper we will obtain precise global stability properties of the one-dimensional steady states of (1.1)-(1.5) for $n = 3$. First, existence, uniqueness, and some properties of the solution of (1.1)-(1.5) are considered in § 2. In § 3, sub- and supersolutions of (1.1)-(1.5), and some properties of the steady state, from the Appendix, are used to show that the (unique) steady state is stable and pointwise globally, asymptotically attracting if $-\infty < m < 0$, while it is unstable if $0 < m < \frac{1}{2}$. Comparison methods do not yield good enough results on the critical adiabatic case $m = 0$, which exhibits infinitely many steady states due to translation invariance. In § 4, a Lyapunov function argument and a nonlinear change of variables will be used to analyze the global stability of the steady states. In particular, we will obtain sufficient conditions on the initial data for the solution of (1.1)-(1.5) to approach the set of steady states as $t \to \infty$, and for it to approach a given steady state.

2. Some preliminary results. In this section we analyze the well-posedness of problem (1.1)-(1.5), as well as some basic properties of its solutions.

The following notation will be used. Let $\Omega \subset \mathbb{R}^n$ be a convex, smooth domain and, for some $T > 0$, let $Q_T = \Omega \times]0, T[$. Let $W_p^q(\Omega)$ (respectively, $W_p^{2q,q}(Q_T)$) be the Sobolev space of those (classes of) functions, $u : \Omega \to \mathbb{R}$ (respectively, $u : Q_T \to \mathbb{R}$) such that $|D^i u|^p$ (respectively, $|D_t^i D_x^j u|^p$) is integrable in Ω (respectively, in Q_T) for all $i \leqq q$ (respectively, for all i and j such that $2i + j \leqq 2q$). The norms of $W_p^q(\Omega)$ and $W_p^{2q,q}(Q_T)$ will be denoted as

$$\|\cdot\|_{p,\Omega}^{(q)} \quad \text{and} \quad \|\cdot\|_{p,Q_T}^{(2q,q)},$$

respectively. $W_{p,\mathrm{loc}}^q(\mathbb{R}^n)$ (respectively, $W_{p,\mathrm{loc}}^{2q,q}(\mathbb{R}^n \times [0, T_0[)$) will be the linear space of those functions $u : \mathbb{R}^n \to \mathbb{R}$ (respectively, $u : \mathbb{R}^n \times [0, T_0[\to \mathbb{R}$) such that $u \in W_p^q(B)$ for all bounded balls $B \subset \mathbb{R}^n$ (respectively, $u \in W_p^{2q,q}(B \times]0, T[)$ for all bounded balls $B \subset \mathbb{R}^n$ and all $T \in]0, T_0[$). For any nonintegral positive number r, $C^r(\bar{\Omega})$ (respectively, $C^{r,r/2}(\bar{Q}_T)$) will be the Hölder space of those functions $u : \Omega \to \mathbb{R}$ (respectively, $u : Q_T \to \mathbb{R}$) having in Ω bounded, uniformly continuous derivatives up to order $[r]$ equal to the integral part of r (respectively, having in Q_T bounded, uniformly continuous derivatives $D_t^i D_x^j u$, for all i and j such that $2i + j < r$) and such that the $[r]$-derivative is uniformly Hölder continuous of order $r - [r]$ in Ω (respectively, the derivatives $D_t^i D_x^j u$ are uniformly Hölder continuous, of order $r - [r]$, in the x variable if $2i + j = [r]$,

and of order $(r-2i-j)/2$ in the t variable if $r-2<2i+j<r$). The norms of $C^r(\bar{\Omega})$ and $C^{r,r/2}(\bar{Q}_T)$ (see, e.g., [24] for their precise definition) will be denoted as

$$|\cdot|_{\Omega}^{(r)} \quad \text{and} \quad |\cdot|_{Q_T}^{(r,r/2)},$$

respectively. Finally, $C^r(\mathbb{R}^n)$ (respectively, $C^{r,r/2}(\mathbb{R}^n \times [0, T_0[$) will be the linear space of those functions $u: \mathbb{R}^n \to \mathbb{R}$ (respectively, $u: \mathbb{R}^n \times [0, T_0[\to \mathbb{R}$) such that $u \in C^r(\bar{B})$ for all bounded balls $B \subset \mathbb{R}^n$ (respectively, $u \in C^{r,r/2}(\bar{B} \times [0, T])$ for all bounded balls $B \subset \mathbb{R}^n$ and all $T \in]0, T_0[$).

We first show that (1.1)-(1.5) possesses a unique classical solution in $0 \leqq t < T_0$, if $-\infty < m < 1$, with $T_0 = \infty$ if $m \leqq 0$.

THEOREM 2.1. *If $-\infty < m < 1$, let $r > 0$ be a noninteger. If $\varphi \in C^{2+r}(\mathbb{R}^n)$ satisfies* (1.2)-(1.4), *then* (1.1)-(1.5) *possess a unique classical solution u in $\mathbb{R}^n \times [0, T_0[$, where $T_0 = \infty$ if $m \leqq 0$ and $T_0 = [2e(1-m)/m] \exp(-a_0)$ if $0 < m < 1$, with*

$$a_0 = \sup \{x_1 - \varphi(x): x \in \mathbb{R}^n\}.$$

Furthermore, $u \in C^{2+r,1+r/2}(\mathbb{R}^n \times [0, T_0[)$ and is such that

$$\sup \{0, x_1 - a(t)\} \leqq u(x, t) \leqq U(x, t) \quad \text{for all } (x, t) \in \mathbb{R}^n \times [0, T_0[, \tag{2.1}$$

where a is given by

$$2e(1-m)\, da/dt = \exp(ma), \qquad a(0) = a_0,$$

and $U > 0$ is the unique solution of

$$\partial U/\partial t = \Delta U \quad \text{in } \mathbb{R}^n \times [0, \infty[, \qquad U(\cdot, 0) = \varphi \quad \text{in } \mathbb{R}^n, \tag{2.2}$$

which satisfies (1.2)-(1.4).

Proof. The solution of (1.1)-(1.5) will be obtained as the limit of the sequence $\{u_k\}$ defined inductively by

$$\partial u_k/\partial t - \Delta u_k + (u_k/2) \exp(mx_1) = (u_{k-1}/2)[1 - \exp(-u_{k-1})] \exp(mx_1), \tag{2.3}$$

$$u_k(x, 0) = \varphi(x), \tag{2.4}$$

where $u_0 = U$ is given by (2.2) and each u_k satisfies (1.2)-(1.4). The coefficient of u_k in (2.3) is unbounded but positive. Therefore, the linear problem (2.3), (2.4) is dissipative, and each u_k is well defined with $u_k \in C^{2+r,1+r/2}(\mathbb{R}^n \times [0, \infty[)$. This is proven by using the estimates of Eidel'man [25, Thm. 3.1, p. 131] for the fundamental solution of (2.3), in standard proofs of the solvability of the Cauchy problem for linear parabolic equations (e.g., in the proof of Theorem 6.1 [24, p. 324]).

The sequence $\{u_k\}$ satisfies, for each $k \geqq 0$,

$$0 \leqq u_{k+1} \leqq u_k \quad \text{in } \mathbb{R}^n \times [0, \infty[, \tag{2.5}$$

as is seen inductively by means of the Phragmèn-Lindelöf (Ph-L) maximum principle [26], [27], when we take into account that the function $u \to u[1 - \exp(-u)]$, appearing in the right-hand side of (2.3), is strictly increasing for $0 \leqq u < \infty$. Then the bounded, monotone sequence $\{u_k\}$ is pointwise convergent to a function u such that

$$0 \leqq u \leqq U \quad \text{in } \mathbb{R}^n \times [0, \infty[, \tag{2.6}$$

as it comes from (2.5).

Let us see that $u \in C^{2+r,1+r/2}(\mathbb{R}^n \times [0, \infty[)$, and that u is a classical solution of (1.1)-(1.5). For each bounded, open ball $B \subset \mathbb{R}^n$, let B' be another ball such that $\bar{B} \subset B'$. Local estimates of the solution of (2.3), (2.4) on $W_p^{2,1}$ and $C^{2+s,1+s/2}$ [24, pp. 355, 352]

imply that, for each $T>0$, each integer $p\geqq 1$ and each noninteger $s\in[0, r]$, there exist constants, $c_1, \cdots, c_4$, depending only on B, B', T, p, and s, such that

$$\|u_j-u_i\|^{(2,1)}_{p,B\times]0,T[}\leqq c_1\|f_j-f_i\|^{(0,0)}_{p,B'\times]0,T[}+c_2\|u_j-u_i\|^{(0,0)}_{p,B'\times]0,T[}, \tag{2.7}$$

$$|u_j-u_i|^{(2+s,1+s/2)}_{B\times]0,T[}\leqq c_3|f_j-f_i|^{(s,s/2)}_{B'\times]0,T[}+c_4|u_j-u_i|^{(s,s/2)}_{B'\times]0,T[} \tag{2.8}$$

for all integers $i, j\geqq 1$, where $f_k=(u_{k-1}/2)[1-\exp(-u_{k-1})]\exp(mx_1)$. Since $u_k\to u$ in $W^{0,0}_p(B'\times]0, T[)=L_p(B'\times]0, T[)$ (monotone convergence theorem [28]) then $f_k\to f=(u/2)[1-\exp(-u)]\exp(mx_1)$, and $\{u_k\}$ and $\{f_k\}$ are Cauchy sequences in the same space. Then $\{u_k\}$ is a Cauchy sequence in $W^{2,1}_p(B\times]0, T[)$ (by (2.7)) and thus it converges (to u) in the same space. Now, if we take $p>(n+2)/(2-r+[r])$, embedding theorems [24, p. 80] imply that $u_k\to u$ in $C^{\alpha,\alpha/2}(\bar{B}\times[0, T])$, where $\alpha=r-[r]$. Estimate (2.8) with $s=\alpha$ implies, by the same argument as above, that $u_k\to u$ in $C^{2+\alpha,1+\alpha/2}(\bar{B}\times[0, T])$ and, by repeating the argument if necessary (i.e., if $r>\alpha$), (2.8) implies that $u_k\to u$ in $C^{2+r,1+r/2}(\bar{B}\times[0, T])$. Then $u\in C^{2+r,1+r/2}(\bar{B}\times[0, T])$ for every bounded ball and every $T>0$ as stated, and u satisfies (1.1) and (1.5), as we see when taking limits in (2.3), (2.4).

We now show that u satisfies (2.1) and, therefore, that it satisfies (1.2)-(1.4). It is enough to prove that $x_1-a(t)\leqq u(x, t)$ for all $(x, t)\in\mathbb{R}^n\times[0, T_0[$ (see (2.6)); this is true since, for all $k\geqq 1$,

$$x_1-a(t)\leqq u_k(x, t)\quad \text{for all } (x, t)\in\mathbb{R}^n\times[0, T_0[,$$

as is seen inductively when the Ph-L maximum principle is applied to $u_k(x, t)-x_1+a(t)$, and it is taken into account that $w(x, t)=x_1-a(t)$ satisfies

$$\partial w/\partial t\leqq\Delta w-\max\{0, (w/2)\exp(mx_1-w)\},\qquad w(x, 0)\leqq\varphi(x)$$

for all $(x, t)\in\mathbb{R}^n\times[0, T_0[$, as is easily seen.

Finally, we see that u is the unique solution of (1.1)-(1.5) in $\mathbb{R}^n\times[0, T_0[$. To this end, first observe that any other solution of (1.1)-(1.5), u', is such that $u'\leqq u_k$ in $\mathbb{R}^n\times[0, T_0[$, for all $k\geqq 0$, as is seen inductively by means of the Ph-L maximum principle. Therefore,

$$u'\leqq u\quad \text{in } \mathbb{R}^n\times[0, T_0[, \tag{2.9}$$

and $W=u-u'$ satisfies

$$\begin{aligned}&\partial W/\partial t-\Delta W=(W/2)(\xi-1)\exp(mx_1-\xi)\quad \text{in } \mathbb{R}^n\times[0, T_0[,\\ &W(x, 0)=0\quad \text{in } \mathbb{R}^n\end{aligned} \tag{2.10}$$

for some function $\xi:\mathbb{R}^n\times[0, T_0[\to\mathbb{R}$ such that $u'\leqq\xi\leqq u$ in $\mathbb{R}^n\times[0, T_0[$. Then the Ph-L maximum principle implies that $W\leqq 0$ in $\mathbb{R}^n\times[0, T_0[$ (observe that the coefficient of W in (2.10) is bounded above, since u and u' satisfy (1.2), (1.3)), and (see (2.9)) the conclusion follows.

Remarks 2.2. Some remarks about Theorem 2.1 are in order:

(A) The function u in the proof of Theorem 2.1 satisfies (1.1), (1.2), (1.4), (1.5) for all $(x, t)\in\mathbb{R}^n\times[0, \infty[$ if $-\infty<m<1$, but if $0<m<1$ we have proved only that u satisfies (1.3) for $0\leqq t<T_0$; for $t\geqq T_0$, u is a maximal (and not necessarily the unique) solution of (1.1), (1.2), (1.4), (1.5). It seems that this result cannot be improved significantly for arbitrary initial data. In fact, some numerical and asymptotic results (see [29]) suggest that for ($0<m<1$ and) appropriate initial data, the maximal solution of (1.1), (1.2), (1.4), (1.5), u, is such that $u(x, t)\to 0$ as $x_1\to+\infty$, uniformly in $(x_2, \cdots, x_n)\in\mathbb{R}^{n-1}$, if $t>T_1$, for some finite T_1; for such initial data, (1.1)-(1.5) cannot

have a solution for all $t \geqq 0$. On the other hand, the proof of Theorem 2.1 is easily extended to show that (1.1)-(1.5) possesses a unique solution in $\mathbb{R}^n \times [0, \infty[$ if the initial datum is such that $\varphi \geqq w(\cdot, 0)$ in $\mathbb{R}^n$, where $w \in C^{2+\alpha, 1+\alpha/2}(\mathbb{R}^n \times [0, \infty[)$ for some $\alpha > 0$ and

$$\partial w/\partial t \leqq \Delta w - (w/2) \exp(mx_1 - w) \quad \text{in } \mathbb{R}^n \times [0, \infty[.$$

(B) The conclusion of Theorem 2.1 (existence and uniqueness of the solution of (1.1)-(1.5) in $\mathbb{R}^n \times [0, T_0[$, for some $T_0 \leqq \infty$) remains true if the nonlinearity $u \exp(mx_1 - u)$ is replaced by a more general one, of the type

$$g(x_1) f(u),$$

where $g : \mathbb{R} \to \mathbb{R}$ and $f : [0, \infty[\to \mathbb{R}$ are positive C^1-functions and:

(i) $f(0) = 0$, $f'(u)$ is bounded in $0 \leqq u < \infty$.

(ii) $g(\xi) f(\xi + c) \to 0$ as $\xi \to \infty$, for any fixed $c \in \mathbb{R}$.

(iii) The boundary condition (1.2) at $x_1 = -\infty$ is replaced by "u bounded" if $g(x_1) \to 0$ (as $x_1 \to -\infty$), and "$u \to 0$" otherwise.

In particular, conditions (i) and (ii) are fulfilled by the first nonlinearity in (1.6) if $1 \leqq p < q$, and by the second if $a > 0$, $p \geqq 1$, and $-\infty < m < 1$.

(C) If, for $m \leqq 0$, the boundary condition (1.2) at $x_1 = -\infty$ is replaced by $u \leqq c$ as $x_1 \to -\infty$, uniformly in $x_2, \cdots, x_n$ and t, for some constant c such that $0 < c < 1$, then the conclusion of Theorem 2.1 remains true, as is easily seen. This fact will be used in § 3, where we will take a supersolution w_2 of (1.1), such that $\lim w_2(x) = \frac{1}{2}$ as $x_1 \to -\infty$, as initial datum.

Problem (1.1)-(1.5) defines a monotone flow, as shown in the following.

THEOREM 2.3. *Under the hypothesis of Theorem* 2.1, *if* u_1 *and* u_2 *are two solutions of* (1.1)-(1.5), *defined in* $0 \leqq t < T_0$, *such that* $u_1(\cdot, 0) \leqq u_2(\cdot, 0)$ *in* $\mathbb{R}^n$, *then* $u_1(\cdot, t) \leqq u_2(\cdot, t)$ *in* $\mathbb{R}^n$, *for all* $t \in [0, T_0[$.

Proof. The monotone sequences that (in the proof of Theorem 2.1) define u_1, u_2, $\{u_{1k}\}$, and $\{u_{2k}\}$ satisfy $u_{1k} \leqq u_{2k}$ in $\mathbb{R}^n \times [0, T_0[$ for all $k \geqq 0$, as is seen inductively by means of the Ph-L maximum principle. Thus we have the conclusion.

As is usual in the literature, a function $w \in C^{2,1}(\mathbb{R}^n \times [0, T_0[)$ is said to be a *supersolution* (respectively, a *subsolution*) of (1.1) in $0 \leqq t < T_0$ if $\partial w/\partial t \geqq \Delta w - (w/2) \exp(mx_1 - w)$ (respectively, $\partial w/\partial t \leqq \Delta w - (w/2) \exp(mx_1 - w)$) in $\mathbb{R}^n \times [0, T_0[$. A sub- or supersolution of (1.1) is said to be *steady* if it does not depend on time.

THEOREM 2.4. *Under the hypothesis of Theorem* 2.1, *if* $w \geqq 0$ *is a supersolution* (*respectively, a subsolution*) *of* (1.1) *in* $0 \leqq t < T_0$ *that satisfies* (1.2)-(1.4), *and if* u *is a solution of* (1.1)-(1.5), *defined in* $0 \leqq t < T_0$, *such that* $u(\cdot, 0) \leqq w(\cdot, 0)$ (*respectively,* $u(\cdot, 0) \geqq w(\cdot, 0)$) *in* $\mathbb{R}^n$, $u(\cdot, t) \leqq w(\cdot, t)$ (*respectively,* $u(\cdot, t) \geqq w(\cdot, t)$) *in* $\mathbb{R}^n$, *for all* $t \in [0, T_0[$.

Proof. If w is a supersolution (respectively, a subsolution) of (1.1), we define the sequence $\{u_k\}$, given by (2.3), (2.4), with $u_0 = w$. As in the proof of Theorem 2.1, it is seen that $0 \leqq u_{k+1} \leqq u_k \leqq w$ (respectively, $w \leqq u_k \leqq u_{k+1} \leqq U$) in $\mathbb{R}^n \times [0, T_0[$ for all $k \geqq 1$, and that $u_k \to u$ as $k \to \infty$; then the conclusion readily follows.

THEOREM 2.5. *Under the hypothesis of Theorem* 2.1, *if* $u(\cdot, 0) = \varphi : \mathbb{R}^n \to \mathbb{R}$ *is a steady supersolution* (*respectively, subsolution*) *of* (1.1), *and satisfies* (1.2)-(1.4), *then* $\partial u/\partial t \leqq 0$ (*respectively,* $\partial u/\partial t \geqq 0$) *in* $\mathbb{R}^n \times [0, T_0[$.

Proof. We consider only the case in which φ is a supersolution. Theorem 2.4 yields: $u(x, t) \leqq \varphi(x) = u(x, 0)$ for all $(x, t) \in \mathbb{R}^n \times [0, T_0[$. Then, for each constant $h \in]0, T_0[$, $w(x, t) = u(x, t + h)$ is a solution of (1.1)-(1.4) such that $w(\cdot, 0) = u(\cdot, h) \leqq u(\cdot, 0)$ in $\mathbb{R}^n$. Thus, Theorem 2.3 leads to $u(\cdot, t + h) = w(\cdot, t) \leqq u(\cdot, t)$ in $\mathbb{R}^n$, for all

$t \in [0, T_0 - h]$. Therefore, for each fixed $x \in \mathbb{R}^n$, the function $t \to u(x, t)$ is nonincreasing and $\partial u/\partial t \leqq 0$ as stated.

Remarks 2.6. (A) Theorems 2.3-2.5 stand when the nonlinearity of (1.1) is modified as in Remark 2.2B, and also, if $m \leqq 0$, when the boundary condition (1.2) at $x_1 = -\infty$ is modified as in Remark 2.2C.

(B) Theorems 2.3-2.5 give properties of the solution of (1.1)-(1.5) that are well known for scalar parabolic equations in bounded domains (see, e.g., [30]).

3. Global stability results in the nonadiabatic case ($m \neq 0$). In this section we analyze global stability properties of the (spatially one-dimensional) steady state of (1.1)-(1.4) under (spatially) three-dimensional perturbations, for $-\infty < m < \frac{1}{2}$, $m \neq 0$. Among the many different definitions of stability, we select the following [31]. Let X be the set of functions $u \in C^{2+r}(\mathbb{R}^3)$, for some $r > 0$, that satisfy (1.2)-(1.4), and let Σ be a family of subsets of X. A steady state u_s of (1.1)-(1.4), such that $u_s \in S$ for all $S \in \Sigma$, will be called Σ-*stable* if for any $S \in \Sigma$ there exists $S' \in \Sigma$ such that $u(\cdot, 0) \in S'$ implies that $u(\cdot, t) \in S$ for all $t > 0$; u_s will be said to be Σ-*unstable* if it is not Σ-stable. Below u_s will be a steady state that depends only on the x_1 coordinate, and the family Σ will be

$$\Sigma = \{S_{\alpha,\beta} : \alpha, \beta > 0\}, \tag{3.1}$$

where

$$S_{\alpha,\beta} = \{u \in X : u_s(x_1 - \alpha) < u(x) < u_s(x_1 + \beta), \text{ for all } x \in \mathbb{R}^3\}.$$

Observe that if $\alpha, \beta > 0$, then $u_s(x_1 - \alpha) < u_s(x_1 + \beta)$ for all $x_1 \in \mathbb{R}$ (Theorems A.4 and A.8 of the Appendix), and $S_{\alpha,\beta}$ is a nonempty open neighborhood of u_s in X with the order topology (i.e., the topology generated by the order intervals of the form $]u_1, u_2[= \{u \in X : u_1(x) < u(x) < u_2(x) \text{ for all } x \in \mathbb{R}^3\}$) defined for $u_1, u_2 \in X$; see [31].

In connection with asymptotic stability, u_s will be said to be *globally pointwise attracting* if $u(\cdot, 0) \in X$ implies that $u(\cdot, t) \to u_s$ pointwise as $t \to \infty$.

We first consider the case $m < 0$.

THEOREM 3.1. *If $m < 0$, then* (1.1)-(1.4) *possesses a unique, spatially one-dimensional steady state u_s that is Σ-stable* (Σ *defined by* (3.1)), *and globally pointwise attracting.*

Proof. We first show that (1.1)-(1.4) has a unique spatially one-dimensional steady state that is globally pointwise attracting. To this end, let us consider the functions $w_1, w_2 : \mathbb{R} \to \mathbb{R}$, defined by

$$w_1(y) = \begin{cases} 0 & \text{if } y \leqq a, \\ A(y-a)^3/64 & \text{if } a < y \leqq a+4, \\ \bar{w}_1(y) & \text{if } a+4 < y, \end{cases}$$

where $A = 2(\sqrt{2}-1)/3$ and $\bar{w}_1$ is the unique solution of

$$d^2\bar{w}_1/dy^2 = (\tfrac{3}{8})\bar{w}_1 \exp(A - \bar{w}_1), \quad \bar{w}_1(a+4) = A, \quad d\bar{w}_1(a+4)/dy = 3A/4,$$

$$w_2(y) = \begin{cases} \frac{1}{2} & \text{if } y \leqq -b, \\ \frac{1}{2} + (y+b)^3[1 - (y+b)/2] & \text{if } -b < y \leqq -b+1, \\ y + b & \text{if } -b+1 < y. \end{cases} \tag{3.2}$$

It is easily seen that w_1 satisfies (1.2), (1.3), w_2 satisfies the boundary conditions considered in Remark 2.2C of § 2, $w_1, w_2 \in C^{2+r}(\mathbb{R})$ for every $r \in]0, 1[$, w_1 is a steady subsolution of (1.1) if $a \geqq |m|^{-1} \ln \frac{4}{3}$, and w_2 is a steady supersolution of (1.1) if

$b \geqq 1 + [1 + \ln \frac{8}{3}]/|m|$. Also, for every function $\varphi : \mathbb{R}^3 \to \mathbb{R}$ satisfying (1.2), (1.3) (uniformly for $x_2, x_3 \in \mathbb{R}$), and (1.4) (uniformly for x_1 on bounded intervals of $\mathbb{R}$), we have

$$w_1(x_1) \leqq \varphi(x) \leqq w_2(x_1) \quad \text{for all } x \in \mathbb{R}^3, \tag{3.3}$$

provided that a and b are sufficiently large, as is easily seen.

Now, for $i = 1$ and 2, let $u_i : \mathbb{R} \times [0, \infty[\to \mathbb{R}$ be given by

$$\partial u_i/\partial t = \partial^2 u_i/\partial y^2 - (u_i/2) \exp(my - u_i) \quad \text{in } \mathbb{R} \times [0, \infty[, \tag{3.4}$$

$$u_1 \to 0, \qquad 0 < u_2 \leqq \tfrac{1}{2} \quad \text{as } y \to -\infty \quad \text{for } 0 \leqq t < \infty, \tag{3.5}$$

$$|u_i - y| \text{ bounded} \quad \text{as } y \to \infty, \qquad 0 \leqq t < \infty, \tag{3.6}$$

$$u_i(y, 0) = w_i(y) \quad \text{for } -\infty < y < \infty, \tag{3.7}$$

where conditions (3.5), (3.6) hold uniformly in $0 \leqq t \leqq T$, for all $T \in]0, \infty[$. The functions u_1 and u_2 are uniquely defined by (3.4)-(3.7), and $u_1, u_2 \in C^{2+r, 1+r/2}(\mathbb{R} \times [0, \infty[)$ (see Theorem 2.1 and Remark 2.2C). Furthermore, if the initial datum of (1.1)-(1.5) satisfies (3.3), then

$$u_1 \leqq u \leqq u_2 \quad \text{in } \mathbb{R}^3 \times [0, \infty[,$$

as is seen when Theorem 2.3 is applied. Also, for each $y \in \mathbb{R}$, the functions $t \to u_1(y, t)$ and $t \to u_2(y, t)$ are monotonic (Theorem 2.5), and bounded since

$$w_1 \leqq u_1(\cdot, t) \leqq u_2(\cdot, t) \leqq w_2 \quad \text{in } \mathbb{R} \quad \text{for all } t \geqq 0, \tag{3.8}$$

as seen by means of Theorems 2.3 and 2.5. Then, for $i = 1$ and 2, $u_i(\cdot, t) \to \tilde{u}_i$ pointwise as $t \to \infty$, for a certain function $\tilde{u}_i : \mathbb{R} \to \mathbb{R}$ such that (see (3.8))

$$w_1 \leqq \tilde{u}_1 \leqq \tilde{u}_2 \leqq w_2 \quad \text{in } \mathbb{R}. \tag{3.9}$$

Thus, according to Lemmas A.1 and A.3 of the Appendix, the conclusion will follow if we prove that $\tilde{u}_1$ and $\tilde{u}_2$ are steady states of (1.1), since these two functions satisfy the boundary conditions (A.4) for $\theta = 1$ (see (3.9)).

To prove that, for $i = 1$ and 2, $\tilde{u}_i$ is a steady state of (1.1) (i.e., that it satisfies (A.1)), let $\psi \in C_0^\infty(\mathbb{R})$ (the space of functions of $C^\infty(\mathbb{R})$ with compact support). We multiply (3.4) by ψ, integrate from $-\infty$ to ∞ in the y variable, and integrate by parts twice to obtain

$$\int_{-\infty}^{\infty} \psi(y)[\partial u_i(y, t)/\partial t]\, dy = \int_{-\infty}^{\infty} \psi''(y) u_i(y, t)\, dy - \int_{-\infty}^{\infty} \psi(y) f(u_i(y, t), y)\, dy,$$

where $f(u, y) = (u/2) \exp(my - u)$. We further integrate from zero to T in the t variable and divide by T, to obtain

$$\begin{aligned} \int_{-\infty}^{\infty} \psi(y)\{[u_i(y, t) - u_i(y, 0)]/T\}\, dy &= \int_{-\infty}^{\infty} \psi''(y) \left(\int_0^T u_i(y, t)\, dt/T \right) dy \\ &\quad - \int_{-\infty}^{\infty} \psi(y) \left(\int_0^T f(u_i(y, t), y)\, dt/T \right) dy. \end{aligned} \tag{3.10}$$

But since, for each $y \in \mathbb{R}$, $u_i(y, t) \to \tilde{u}_i(y)$ as $t \to \infty$, we have

$$[u_i(y, t) - u_i(y, 0)]/T \to 0, \qquad \int_0^T u_i(y, t)\, dt/T \to \tilde{u}_i(y),$$

$$\int_0^T f(u_i(y, t), y)\, dt/T \to f(\tilde{u}_i(y), y) \quad \text{pointwise as } T \to \infty. \tag{3.11}$$

Furthermore, the left-hand sides in the limits (3.11) are uniformly bounded in every bounded interval of $\mathbb{R}$ (see (3.8)) and, in particular, in supp ψ. Then if we let $T \to \infty$ in (3.10), the dominated convergence theorem [28] yields

$$\int_{-\infty}^{\infty} \psi''(y)\tilde{u}_i(y)\, dy = \int_{-\infty}^{\infty} \psi(y) f(\tilde{u}_i(y), y)\, dy,$$

for all $\psi \in C_0^\infty(\mathbb{R})$. Therefore $\tilde{u}_i$ satisfies (A.1) as a distribution (observe that $\tilde{u}_i \in L_{2,\mathrm{loc}}(\mathbb{R})$, as we see by means of the dominated convergence theorem when taking into account (3.8)) and, since the function $y \to f(\tilde{u}_i(y), y)$ belongs to $L_{2,\mathrm{loc}}(\mathbb{R})$, $\tilde{u}_i \in W^2_{2,\mathrm{loc}}(\mathbb{R})$. Also, $\tilde{u}_i \in W^p_{2,\mathrm{loc}}(\mathbb{R})$ for all $p > 2$, as is seen by reiterating the argument. Then embedding theorems [28] imply that $\tilde{u}_i \in C^\infty(\mathbb{R})$ and satisfies (A.1) as stated.

Finally, $\tilde{u}_1 = \tilde{u}_2 = u_s$ is Σ-stable, as comes out when Theorems 2.3 and 2.4 are applied, and it is taken into account that, if $\alpha, \beta \geqq 0$, then the functions $x \to u_s(x_1 - \alpha)$ and $x \to u_s(x_1 + \beta)$ are steady sub- and supersolutions of (1.1), respectively, as is easily seen.

COROLLARY 3.2. *If $m < 0$ and $n = 3$, then* (1.1)-(1.4) *has a unique steady state u_s which depends only on the x_1 variable.*

Proof. The steady state of Theorem 3.1 is necessarily the unique steady state of (1.1)-(1.4) since it is globally attracting.

Remarks 3.3. (A) In Theorem 3.1 we have shown that, for every initial datum φ satisfying (1.2)-(1.4), the solution of (1.1)-(1.5) is such that $u(\cdot, t) \to u_s$ pointwise as $t \to \infty$. It may be seen that the convergence is uniform on compact subsets of $\mathbb{R}^3$, but it is not uniform in $\mathbb{R}^3$ for arbitrary initial data. For example, if φ depends only on the x_1 variable, $\varphi(x_1) - x_1$ has a limit as $x_1 \to \infty$, and $\lim(\varphi(x_1) - x_1) \neq \lim(u_s(x_1) - x_1)$ as $x_1 \to \infty$, then the solution of (1.1)-(1.5) satisfies, for each $t > 0$, $\lim(u(x_1, t) - x_1) = \lim(\varphi(x_1) - x_1) \neq \lim(u_s(x_1) - x_1)$ as $x_1 \to \infty$, as may be seen.

(B) Corollary 3.2 shows that, in addition to the (spatially one-dimensional) steady state of (1.1)-(1.4) found by Liñán [2], there are no other steady states, possibly depending on the x_2 and x_3 coordinates. For a more precise information about the (unique) steady state of (1.1)-(1.4), see Theorem A.4 in the Appendix.

We now consider the case $m > 0$, in which (1.1)-(1.4) possesses a unique spatially one-dimensional steady state (see Theorem A.8 in the Appendix), that is expected to be unstable, according to the numerical results by Peters [20].

THEOREM 3.4. *If $0 < m < \frac{1}{2}$ and $n = 3$, let u_s be the (unique) spatially one-dimensional steady state of* (1.1)-(1.4). *Then*:

(A) *If the initial state* (1.5) *satisfies $\varphi(x) \geqq u_s(x_1 + \alpha)$ for some $\alpha > 0$ and all $x \in \mathbb{R}^3$, then the solution of* (1.1)-(1.5) *is uniquely defined for all $t \geqq 0$ and such that, for each $x \in \mathbb{R}^3$, $\lim u(x, t) = \infty$ as $t \to \infty$.*

(B) *If a solution of* (1.1)-(1.5) *is defined for all $t \geqq 0$ and the initial state satisfies $\varphi(x) \leqq u_s(x_1 - \alpha)$ for all $x \in \mathbb{R}^3$ and some $\alpha > 0$, then $\lim u(\cdot, t) = 0$ pointwise as $t \to \infty$.*

(C) *The steady state u_s is Σ-unstable (Σ defined by* (3.1)).

Remark 3.5. Under the hypothesis of Theorem 3.4B, the solution of (1.1)-(1.5) is uniquely defined in $0 \leqq t < \infty$ whenever the initial state satisfies $u(x, 0) \geqq \varphi(x)$ for

all $x \in \mathbb{R}^n$ (φ as given in Theorem 3.4B), as is seen when taking the solution considered in Theorem 3.4B as w in Remark 2.2A. If the solution of Theorem 3.4B is not assumed to exist for all $t \geqq 0$ but the other hypothesis is maintained, then the maximal solution of (1.1), (1.2), (1.4), (1.5) (that exists for $0 \leqq t < \infty$; see Remark 2.2A) satisfies the conclusion, as is seen after slight modifications in the proof.

Proof of Theorem 3.4. (A) To prove that u is uniquely defined for all $t \geqq 0$, observe that $w = u_s$ satisfies the required properties of Remark 2.2A. It is sufficient to prove the remaining parts of the statement when $\varphi(x) = u_s(x_1 + \alpha)$ (Theorem 2.3); then $u(x, t) = u(x_1, t)$ does not depend on the x_2 and x_3 coordinates, and satisfies $\partial u(x_1, t)/\partial t \geqq 0$ for all $(x_1, t) \in \mathbb{R} \times [0, \infty[$ (apply Theorem 2.5 and take into account that the function $x_1 \to u_s(x_1 + \alpha)$ is a subsolution of (1.1) since $\alpha > 0$). To prove that $\lim u(x, t) \to \infty$ pointwise as $t \to \infty$ suppose, on the contrary, that for some finite c, $x_1^0 \in \mathbb{R}$, $u(x_1^0, t) \leqq c$ for all $t > 0$. Then

$$u(x_1, t) \leqq c \quad \text{for all } (x_1, t) \in]-\infty, x_1^0] \times [0, \infty[,$$

$$u(x_1, t) \leqq c + x_1 - x_1^0 \quad \text{for all } (x_1, t) \in [x_1^0, \infty[\times [0, \infty[,$$

as is seen by applying the Ph-L maximum principle on the intervals $]-\infty, x_1^0[$ and $]x_1^0, \infty[$. Then, for each $x_1 \in \mathbb{R}$, the increasing function $t \to u(x_1, t)$ is bounded above and, by the argument of the proof of Theorem 3.1, $u(x_1, t) \to \bar{u}_s(x_1)$ pointwise as $t \to \infty$, where $\bar{u}_s$ is a solution of (A.1), (A.2) such that $u_s(x_1) < \bar{u}_s(x_1)$ for all $x_1 \in \mathbb{R}$. But this is not possible, according to Theorem A.4.

(B) As in the proof of part A, it is sufficient to prove the result when $\varphi(x) = u_s(x_1 - \alpha)$. Then the solution does not depend on the x_2 and x_3 coordinates, but is defined for all $t \geqq 0$ (Remark 3.5), and, by the argument of the proof of part A ($t \to u(x_1, t) \geqq 0$ is now decreasing), $u(x_1, t) \to \bar{u}_s(x_1)$ pointwise as $t \to \infty$, where $\bar{u}_s$ satisfies (A.1) and $0 \leqq \bar{u}_s(x_1) < u_s(x_1)$ for all $x_1 \in \mathbb{R}$. Then $\bar{u}_s(x_1) = 0$ for all $x_1 \in \mathbb{R}$ (Lemma A.10 in the Appendix) and the conclusion follows.

(C) Apply parts A and B above.

4. Global stability results in the adiabatic case ($m = 0$). Let us now consider the critical case $m = 0$. Again, we are interested in the stability properties of the spatially one-dimensional steady states of (1.1)-(1.4) (there are infinitely many due to translation invariance; see Theorem A.2 in the Appendix) under spatially three-dimensional perturbations. The last part of the proof of Theorem 3.1 is readily extended to yield Theorem 4.1 below.

THEOREM 4.1. *If $m = 0$, then every spatially one-dimensional steady state of* (1.1)-(1.4) *is Σ-stable* (Σ as given in (3.1)).

The remaining part of Theorem 3.1 cannot hold in this case, since there is no *unique* steady state now. We can easily be convinced that comparison methods alone cannot lead us further in the analysis of asymptotic stability properties if $m = 0$. Linear stability of the steady states of (1.1)-(1.4) is easily analyzed for $n = 1$. Although linear stability results do not solve the problem, they are enlightening, and help us to avoid the pursuit of ideas that cannot work in this case. For $n = 1$, the linear eigenvalue problem associated with a given steady state u_s of (1.1)-(1.3) is

$$u'' - f'(u_s)u = \omega u \quad \text{in } -\infty < x < \infty, \tag{4.1}$$

where $f(u) = (u/2)\exp(-u)$. The steady state u_s is easily seen to be such that u_s''/u_s', u_s'''/u_s', and $f(u_s)$ are bounded in $-\infty < x < \infty$. We consider (4.1) in $L2(\mathbb{R})$ (where (4.1) is self-adjoint) and in $C(\bar{\mathbb{R}})$ (the space of real, bounded, uniformly continuous

functions on $\mathbb{R}$ with the sup norm). The function $u_s' \in C(\bar{\mathbb{R}})$ satisfies

$$u_s''' - f'(u_s)u_s' = 0 \quad \text{in } -\infty < x < \infty, \tag{4.2}$$

and thus is an eigenfunction of (4.1) associated with $\omega = 0$. Then the general solution of the homogeneous equation (4.1) is easily calculated for $\omega = 0$, and it is seen that $\omega = 0$ is a simple eigenvalue in $C(\bar{\mathbb{R}})$ and that it is not an eigenvalue in $L2(\mathbb{R})$. Also, if $\operatorname{Re}\omega > 0$, then any bounded eigenfunction of (4.1) belongs to $L2(\mathbb{R})$, as is seen from its asymptotic behavior as $x \to \pm\infty$ (see, e.g., [32]); then $\omega \in \mathbb{R}$ also in $C(\bar{\mathbb{R}})$, and any eigenfunction u of (4.1) is such that

$$\omega \int_{-\infty}^{\infty} u^2\,dx = -\int_{-\infty}^{\infty} [u' - uu_s''/u_s']^2\,dx, \tag{4.3}$$

as seen after multiplication of (4.1) by u, integration from $-\infty$ to ∞, substitution of (4.2), and integration by parts. To obtain (4.3) observe that, since $u \in L2(\mathbb{R})$, $u'' \in L2(\mathbb{R})$ (see (4.1)), and $u' \in L2(\mathbb{R})$ as shown by interpolation inequalities (see, e.g., [28, p. 70]). Equation (4.3) implies that every eigenvalue of (4.1) in $C(\bar{\mathbb{R}})$ or in $L2(\mathbb{R})$ is such that $\operatorname{Re}\omega \leqq 0$. If the continuous spectrum of (4.1), σ, were such that $\max \operatorname{Re}\sigma < 0$, then standard results on linear stability [33, p. 108, Exercise 6] would show that if $u(\cdot, 0)$ is in a certain neighborhood (in $C(\bar{\mathbb{R}})$) of u_s, then the solution of (1.1)-(1.5) for $n = 1$ approaches exponentially a translate of u_s as $t \to \infty$. Unfortunately, the continuous spectrum of (4.1), in $L2(\mathbb{R})$ and in $C(\bar{\mathbb{R}})$, is $\sigma =]-\infty, 0] \subset \mathbb{R}$ [33, p. 140], and the result above does not apply. Observe that the spectrum of (4.1) is equal to that of the heat equation (which also has infinitely many steady states in $C(\bar{\mathbb{R}})$), which, as is well known, exhibits erratic behavior as $t \to \infty$ for appropriate initial conditions in every neighborhood (in $C(\bar{\mathbb{R}})$) of each steady state (see, e.g., [26, p. 349]). Finally, let us point out that problem (1.1)-(1.3) for $n = 1$ has some features in common with one-dimensional reaction-diffusion problems exhibiting travelling fronts, which have received considerable attention in the literature (see [33] and [34] and references given therein).

We first consider problem (1.1)-(1.5) in one space dimension. The first part of the following theorem contains an invariant principle that holds for a general class of semilinear parabolic equations *in a bounded domain*, as is well known [33, § 4.3]. There are some more recent extensions of this principle (see, e.g., [35], [36]) that, unfortunately, do not apply to (1.1)-(1.5). Observe also that the result of Theorem 4.2B implies stabilization of certain solutions of (1.1)-(1.5) in a very weak sense, and resembles well-known results for travelling fronts, such as those appearing in the celebrated Kolmogorov-Petrovsky-Piscounov model equation [33, p. 134].

THEOREM 4.2. *If $n = 1$, let the hypothesis of Theorem* 2.1 *be satisfied, and let u_s be a steady state of* (1.1)-(1.3). *If the initial state* (1.5) *is such that $u_s(x-\alpha) \leqq \varphi(x) \leqq u_s(x+\beta)$, $\varphi'(x) > 0$ in $-\infty < x < \infty$, for some finite constants α and β, and $\varphi' - u_s' \in W_2^6(\mathbb{R})$, then the unique solution of* (1.1)-(1.3), (1.5) *is such that*

$$u_s(x-\alpha) \leqq u(x,t) \leqq u_s(x+\beta), \qquad u_x(x,t) \geqq 0 \quad \textit{for all } (x,t) \in \mathbb{R} \times [0,\infty[, \tag{4.4}$$

$$u_x - u_s' \in W_2^{6,3}(\mathbb{R} \times [0,T[) \quad \textit{for all } T \in [0,\infty[, \tag{4.5}$$

and satisfies the following properties:

(A) *There exists a C^2 bounded function $\xi : [0,\infty[\to \mathbb{R}$ such that $u(x,t) - u_s(x+\xi(t)) \to 0$, uniformly on bounded intervals of $\mathbb{R}$, as $t \to \infty$.*

(B) *$\xi'(t) \to 0$ as $t \to \infty$.*

Proof. The first inequalities (4.4) are readily obtained by applying Theorem 4.1. Then (4.5) is obtained by standard estimates on $W_2^{2m,m}$ spaces (see, e.g.,

[24, Chap. IV, § 9]) applied to the (linear) parabolic Cauchy problem for $u_x - u_s'$ which is obtained by differentiating (1.1) with respect to x. Then $\lim (u_x - u_s') = 0$ as $x \to \pm\infty$, uniformly in $0 \leqq t < T$ for all $T \in [0, \infty[$, and the second inequality (4.4) is readily obtained when the Ph-L maximum principle is applied to the equation obtained by differentiating (1.1) with respect to x, and we take into account that $\varphi'(x) \geqq 0$ in $-\infty < x < \infty$. We now prove properties A and B.

(A) We define the energy integral

$$H(t) = \int_{-\infty}^{\infty} [(u_x - u_s')^2 + (1 + u_s - u_s u + u_s^2) \exp(-u_s) - (1+u) \exp(-u)]\, dx,$$

which, when using (4.5), is easily seen to satisfy

$$H'(t) = -2 \int_{-\infty}^{\infty} u_t^2\, dx.$$

Then the function $t \to H(t)$ is monotonically decreasing; since it is bounded below (see (4.4)), it has a limit as $t \to \infty$, and

$$\text{(4.6)} \qquad \int_{-\infty}^{\infty} (u_x - u_s')^2\, dx \quad \text{and} \quad \int_0^t dt \int_{-\infty}^{\infty} u_t^2\, dx \quad \text{are bounded in } 0 \leqq t < \infty.$$

On the other hand, when differentiating (1.1) with respect to t, multiplying by u_t, integrating in the x variable from $-\infty$ to ∞, and taking into account (4.4), (4.5) we obtain

$$\text{(4.7)} \qquad \frac{1}{2} \frac{d}{dt} \int_{-\infty}^{\infty} u_t^2\, dx \leqq -\int_{-\infty}^{\infty} u_{tx}^2\, dx + k \int_{-\infty}^{\infty} u_t^2\, dx \quad \text{in } 0 \leqq t < \infty$$

for a certain positive, finite constant k. When we take into account (4.6), this inequality yields

$$\text{(4.8)} \qquad \frac{d}{dt} \int_{-\infty}^{\infty} u_t^2\, dx \text{ is bounded above} \quad \text{in } 0 \leqq t < \infty.$$

Then, (4.6) and (4.8) imply that

$$\text{(4.9)} \qquad \int_{-\infty}^{\infty} u_t^2\, dx \to 0 \quad \text{as } t \to \infty.$$

Now, when using Hölder's inequality, (4.6) and (4.9) yield

$$0 \leqq \left| \int_{-\infty}^{x} u_x u_t\, dx \right| \leqq \left[\int_{-\infty}^{\infty} (u_x - u_s')^2\, dx \int_{-\infty}^{\infty} u_t^2\, dx \right]^{1/2} + \left[\int_{-\infty}^{x} u_s'^2\, dx \int_{-\infty}^{\infty} u_t^2\, dx \right]^{1/2} \to 0 \quad \text{as } t \to \infty,$$

uniformly on each bounded interval of $\mathbb{R}$. Then, when multiplying (1.1) by u_x and integrating in the x variable from $-\infty$ to x, we easily obtain

$$\text{(4.10)} \qquad u_x^2 - 1 + (1+u) \exp(-u) \to 0 \quad \text{as } t \to \infty,$$

uniformly on each bounded interval of $\mathbb{R}$.

Finally, for each $t > 0$, let us define $\xi(t)$ as the unique solution of the equation

$$\text{(4.11)} \qquad u(0, t) = u_s(\xi(t)).$$

Since $u_s'(x) > 0$ in $-\infty < x < \infty$, $t \to \xi(t)$ is a well-defined C^2-function in $0 \leqq t < \infty$ (Inverse Function Theorem) and (see (4.4))

$$-\infty < -\alpha \leqq \xi(t) \leqq \beta < \infty \quad \text{in } 0 \leqq t < \infty. \tag{4.12}$$

Then, since u satisfies (4.10), (4.11), $u_x(x, t) \geqq 0$ for all $(x, t) \in \mathbb{R} \times [0, \infty[$ (see (4.4)) and, for each fixed $t \geqq 0$, the function $x \to u_s(x + \xi(t))$ satisfies (A.6), standard results on continuous dependence on parameters of the solution of the Cauchy problem for ordinary differential equations [32] imply that $u(x, t) - u_s(x + \xi(t)) \to 0$, uniformly on bounded intervals of $\mathbb{R}$, as $t \to \infty$.

(B) We first observe that

$$\int_{-\infty}^{\infty} u_{xx}^2 \, dx \text{ is bounded} \quad \text{in } 0 \leqq t < \infty,$$

as obtained from (1.1) when taking into account (4.4) and (4.9). Then for each $(x, t) \in \mathbb{R} \times [0, \infty[$ we have

$$[u_x(x, t) - u_s'(x)]^2 = 2 \int_{-\infty}^{x} (u_x - u_s')(u_{xx} - u_s'') \, dx$$

$$\leqq 2 \left[\int_{-\infty}^{\infty} (u_x - u_s')^2 \, dx \int_{-\infty}^{\infty} (u_{xx} - u_s'')^2 \, dx \right]^{1/2},$$

and (see (4.6))

$$u_x(x, t) \text{ is uniformly bounded} \quad \text{in } \mathbb{R} \times [0, \infty[. \tag{4.13}$$

On the other hand, when integrating (4.7) from zero to ∞ and taking into account (4.6) and (4.9), we obtain

$$\int_0^t dt \int_{-\infty}^{\infty} u_{xt}^2 \, dx \text{ is bounded} \quad \text{in } 0 \leqq t < \infty. \tag{4.14}$$

In addition, when differentiating (1.1) twice with respect to x and to t, multiplying by u_{xt}, integrating in the x variable from $-\infty$ to ∞, and taking into account (4.4), (4.5), and (4.13), we obtain

$$\frac{1}{2} \frac{d}{dt} \int_{-\infty}^{\infty} u_{xt}^2 \, dx \leqq k_1 \int_{-\infty}^{\infty} u_{xt}^2 \, dx + k_2 \left[\int_{-\infty}^{\infty} u_{xt}^2 \, dx \int_{-\infty}^{\infty} u_t^2 \, dx \right]^{1/2}$$

for certain finite constants k_1 and k_2. If we integrate this inequality from zero to t, and take into account (4.6) and (4.14), we get that

$$\int_{-\infty}^{\infty} u_{xt}^2 \, dx \text{ is bounded} \quad \text{in } 0 \leqq t < \infty.$$

Then (4.9) yields

$$0 \leqq u_t(0, t)^2 = 2 \int_{-\infty}^{0} u_t u_{xt} \, dx \leqq 2 \left[\int_{-\infty}^{\infty} u_t^2 \, dx \int_{-\infty}^{\infty} u_{tx}^2 \, dx \right]^{1/2} \to 0 \quad \text{as } t \to \infty,$$

and, since $u_t(0, t) = u_s'(\xi(t))\xi'(t)$ and $\xi(t)$ satisfies (4.12), the conclusion follows.

Observe that Theorem 4.2 does not imply that u approaches a steady state of (1.1)-(1.3) as $t \to \infty$. Nevertheless, if $u(x, 0) - u_s(x) \to 0$ as $x \to +\infty$, for a certain steady state u_s, then u approaches u_s as $t \to \infty$, as is proven below. To this end, let us assume that the hypotheses of Theorem 4.2 are satisfied (less than that is needed in the following

analysis, but more generality will not be necessary and is avoided for the sake of brevity). Let us introduce the function $\rho : \mathbb{R} \times [0, \infty[\to \mathbb{R}$ by

$$u_s(\rho(x, t)) = u(x, t).$$

Since $u_s'(x) > 0$ in $-\infty < x < \infty$, ρ is a well-defined function of $C^{r, r/2}(\mathbb{R} \times [0, \infty[)$ for some $r > 3$ (Inverse Function Theorem) and satisfies

$$\partial \rho / \partial t = \partial^2 \rho / \partial x^2 + g(\rho)[(\partial \rho / \partial x)^2 - 1] \quad \text{in } \mathbb{R} \times [0, \infty[, \tag{4.15}$$

as is easily seen, where

$$g(\rho) = u_s''(\rho)/u_s'(\rho)$$

is positive and uniformly bounded and

$$g'(\rho) = [u_s'(\rho)u_s'''(\rho) - u_s''(\rho)^2]/u_s'(\rho)^2$$

is uniformly bounded. To prove that, take into account that u_s satisfies (A.1) and (A.6).

In addition, the function ρ satisfies

$$x - \alpha \leqq \rho(x, t) \leqq x + \beta, \qquad \rho_x(x, t) \geqq 0 \quad \text{in } \mathbb{R} \times [0, \infty[, \tag{4.16}$$

from (4.4).

The required result will be easily obtained from the following two lemmas.

LEMMA 4.3. *Under the assumptions above, if* $0 \leqq \rho_x(x, 0) \leqq 1$ *(respectively,* $1 \leqq \rho_x(x, 0) < \infty$*) in* $-\infty < x < \infty$*, then* $0 \leqq \rho_x(x, t) \leqq 1$ *(respectively,* $1 \leqq \rho_x(x, t) < \infty$*) for all* $(x, t) \in \mathbb{R} \times [0, \infty[$.

Proof. Let us first show that there is a finite constant k such that

$$0 \leqq \rho_x(x, t) \leqq k \quad \text{for all } (x, t) \in \mathbb{R} \times [0, \infty[. \tag{4.17}$$

ρ_x is nonnegative (see (4.16)) and, since $\rho_x(x, 0)$ is bounded in $-\infty < x < \infty$, $u_s'(x) \exp(-x/\sqrt{2})$ has a limit as $x \to -\infty$ (as is seen from the asymptotic behavior of (A.1), (A.2)), and ρ satisfies (4.16), we have

$$0 \leqq u_x(x, 0) = u_s'(\rho(x, 0))\rho_x(x, 0) \leqq k_1 \exp(x/\sqrt{2}) \quad \text{in } -\infty < x < \infty \tag{4.18}$$

for a certain finite constant k_1. Then we can see that

$$u_x(x, t) \leqq k_2 \exp(x/\sqrt{2}) \quad \text{for all } (x, t) \in]-\infty, x_0[\times [0, \infty[, \tag{4.19}$$

where x_0 is any point of $\mathbb{R}$ such that $u_s(x_0 + \beta) < 1$ (then $u(x, t) \leqq 1$ in $]-\infty, x_0] \times [0, \infty[$; see (4.16)) and $k_2 = \max\{k_1, \sup\{u_x(x_0, t) : t \geqq 0\}\}$ (k_2 is finite; see (4.13)). To prove that (4.19) holds, apply the Ph-L maximum principle in $-\infty < x \leqq x_0$ to the equation obtained when (1.1) is differentiated with respect to x, and take into account (4.13) and (4.18). Then $\rho_x(x, t) = u_x(x, t)/u_s'(\rho(x, t))$ satisfies (4.17) since (i) $u_s'(x) \exp(-x/\sqrt{2})$ and $u_s'(x)$ are bounded below by a strictly positive constant in $]-\infty, x_0]$ and in $[x_0, \infty[$, respectively; (ii) the function $x \to u_s'(x)$ is strictly increasing in $-\infty < x < \infty$; (iii) ρ satisfies (4.16); and (iv) u_x satisfies (4.13) and (4.19).

Then the conclusion of the lemma readily follows when the Ph-L maximum principle is applied to the equation obtained when (4.15) is differentiated with respect to x, and it is taken into account that g and g' are uniformly bounded, and that (4.17) holds.

LEMMA 4.4. *If, in addition to the assumptions of Theorem* 4.2, *the initial condition* (1.5) *is such that* $\varphi(x) = u_s(x)$ *in* $k < x < \infty$, *for some finite constant* k, *then* $u(x, t) \to u_s(x)$ *pointwise as* $t \to \infty$.

Proof. Let the function $\rho^0:\mathbb{R}\to\mathbb{R}$ be defined by $u_s(\rho^0(x))=\varphi(x)$; as above, ρ^0 is a well-defined function such that $x-\alpha\leqq\rho^0(x)\leqq x+\beta$ in $-\infty<x<\infty$.

It is easily seen also that there exist two functions, ρ_1^0, $\rho_2^0\in C^\infty(\mathbb{R})$, such that

$$x-\alpha\leqq\rho_1^0(x)\leqq\rho^0(x)\leqq\rho_2^0(x)\leqq x+\beta,\qquad -\infty<x<\infty, \tag{4.20}$$

$$\rho_1^0(x)=x-\alpha,\qquad \rho_2^0(x)=x+\beta\qquad \text{in } -\infty<x<k_1, \tag{4.21}$$

$$\rho_1^0(x)=\rho_2^0(x)=\rho^0(x)=x\qquad \text{in } k_2<x<\infty, \tag{4.22}$$

$$1\leqq d\rho_1^0/dx<\infty,\qquad 0\leqq d\rho_2^0/dx\leqq 1\quad \text{in } -\infty<x<\infty \tag{4.23}$$

for some finite constants k_1 and k_2. Then the assumptions of Theorem 4.2 are satisfied for the functions u_1 and u_2 given by (1.1)-(1.3) and

$$u_i(x,0)=u_s(\rho_i^0(x))\quad \text{in } -\infty<x<\infty\quad \text{for } i=1,2.$$

Furthermore, since $u_1(\cdot,0)\leqq u(\cdot,0)\leqq u_2(\cdot,0)$ and $u_1(\cdot,0)\leqq u_s\leqq u_2(\cdot,0)$ in $\mathbb{R}$ (see (4.20)-(4.23)), Theorem 2.3 yields

$$u_1(\cdot,t)\leqq u(\cdot,t)\leqq u_2(\cdot,t),\qquad u_1(\cdot,t)\leqq u_s\leqq u_2(\cdot,t)\quad \text{in } \mathbb{R} \tag{4.24}$$

for all $t\geqq 0$. Then the conclusion follows if we prove that

$$u_i(x,t)\to u_s(x)\quad \text{pointwise as } t\to\infty. \tag{4.25}$$

To this end, let us define, for $i=1$ and 2, the function $\rho_i:\mathbb{R}\times[0,\infty[\to\mathbb{R}$ by $u_i(x,t)=u_s(\rho_i(x,t))$; that function is again well defined and $\rho_i(x,0)=\rho_i^0(x)$ in $-\infty<x<\infty$. Then (apply Lemma 4.3 and take into account (4.23)),

$$1\leqq\partial\rho_1/\partial x<\infty,\qquad 0\leqq\partial\rho_2/\partial x\leqq 1\quad \text{in } \mathbb{R}\times[0,\infty[, \tag{4.26}$$

and, since ρ_i satisfies (4.15) and $\partial u_i(x,t)/\partial t=u_s'(\rho_i(x,t))\partial\rho_i/\partial t$, we have, for $i=1$ and 2 and for all $t\geqq 0$,

$$\begin{aligned}\frac{d}{dt}\int_{-\infty}^{\infty}[u_i(x,t)-u_s(x)]\,dx&=\int_{-\infty}^{\infty}u_s'(\rho_i(x,t))\frac{\partial\rho_i(x,t)}{\partial x}\,dx\\&=\int_{-\infty}^{\infty}u_s'(\rho_i(x,t))\frac{\partial^2\rho_i(x,t)}{\partial x^2}\,dx\\&\quad+\int_{-\infty}^{\infty}u_s''(\rho_i(x,t))\left[\left(\frac{\partial\rho_i(x,t)}{\partial x}\right)^2-1\right]dx\\&=1-\int_{-\infty}^{\infty}u_s''(\rho_i(x,t))\,dx\\&=\int_{-\infty}^{\infty}u_s''(\rho_i(x,t))\left[\frac{\partial\rho_i(x,t)}{\partial x}-1\right]dx,\end{aligned}$$

where the manipulations on the improper integral required to obtain the first equality are easily seen to be justified. The third equality is obtained by integration by parts in the first integral of the left-hand side, when taking into account that $\rho_{ix}(x,t)$ is bounded and $u_s'(\rho_i(x,t))\to 0$ as $x\to-\infty$, and that $\rho_{ix}(x,t)\to 1$ and $u_s'(\rho_i(x,t))\to 1$ as $x\to\infty$, for all $t\geqq 0$; the last equality is obtained when taking into account that

$$\int_{-\infty}^{\infty}u_s''(\rho_i(x,t))\frac{\partial\rho_i(x,t)}{\partial t}\,dx=\int_{-\infty}^{\infty}\frac{d}{dx}[u_s'(\rho_i(x,t))]\,dx=1.$$

Then the functions $t \to \int_{-\infty}^{\infty} (u_s - u_1)\, dx$ and $t \to \int_{-\infty}^{\infty} (u_2 - u_s)\, dx$ are monotonically decreasing (see (4.26)) and nonnegative (see (4.24)) in $0 \leqq t < \infty$.

Therefore, for $i = 1$ and 2,

$$\int_{-\infty}^{\infty} |u_i - u_s|\, dx \text{ is bounded} \quad \text{in } 0 \leqq t < \infty;$$

since, in addition, u_1 and u_2 satisfy property A of Theorem 4.2, (4.25) readily follows, and the proof is complete.

Finally, we prove the main result of this section.

THEOREM 4.5. *If $m = 0$ and $n = 3$, let u_s be a spatially one-dimensional steady state of* (1.1)-(1.3), *and let the assumptions of Theorem* 2.1 *hold. If the initial state* (1.5) *is such that*

$$u_s(x_1 - \alpha) \leqq \varphi(x) \leqq u_s(x_1 + \beta) \quad \textit{for all } x \in \mathbb{R}^3$$

for some finite constants α and β, and

$$\lim (\varphi(x) - u_s(x_1)) = 0 \quad \textit{as } x_1 \to \infty,$$

uniformly for $(x_2, x_3) \in \mathbb{R}^2$, then the solution of (1.1)-(1.5) *is such that* $\lim u(\cdot, t) \to u_s$ *pointwise as $t \to \infty$.*

Proof. From the assumptions above, it is clear that, for each $\varepsilon > 0$, there exist two functions, φ_1^ε, $\varphi_2^\varepsilon \in C^\infty(\mathbb{R})$, that satisfy (1.2), (1.3), and

$$\begin{aligned} &\varphi_1^\varepsilon(x_1) \leqq \varphi(x) \leqq \varphi_2^\varepsilon(x_1) \quad \text{for all } x \in \mathbb{R}^3, \\ &\varphi_1^\varepsilon(x_1) = u_s(x_1 - \alpha), \qquad \varphi_2^\varepsilon(x_1) = u_s(x_1 + \beta) \quad \text{in } -\infty < x_1 < k_1, \\ &\varphi_1^\varepsilon(x_1) = u_s(x_1 - \varepsilon), \qquad \varphi_2^\varepsilon(x_1) = u_s(x_1 + \varepsilon) \quad \text{in } k_2 < x_1 < \infty, \end{aligned} \tag{4.27}$$

for some finite constants k_1 and k_2. For $i = 1$ and 2, let us define the functions $u_i^\varepsilon : \mathbb{R} \times [0, \infty[\to \mathbb{R}$ by (1.1)-(1.3) (with $n = 1$) and

$$u_i^\varepsilon(x_1, 0) = \varphi_i^\varepsilon(x_1) \quad \text{in } -\infty < x_1 < \infty.$$

Then (apply Theorem 2.3 and take into account (4.27))

$$u_1^\varepsilon(x_1, t) \leqq u(x, t) \leqq u_2^\varepsilon(x_1, t) \quad \text{for all } (x, t) \in \mathbb{R}^3 \times [0, \infty[, \tag{4.28}$$

and (apply Lemma 4.4)

$$u_1^\varepsilon(x_1, t) \to u_s(x_1 - \varepsilon), \qquad u_2^\varepsilon(x_1, t) \to u_s(x_1 + \varepsilon) \quad \text{pointwise as } t \to \infty. \tag{4.29}$$

Since (4.28) and (4.29) are true for all $\varepsilon > 0$, the conclusion follows.

Remarks 4.6. Some remarks about the result above are in order.

(A) Theorem 4.5 is true also in one and two space dimensions (after obvious modifications).

(B) The results of this section and, in particular, of Theorem 4.5, stand when the nonlinearity of (1.1) is replaced by a positive C^3-function $f : [0, \infty[\to \mathbb{R}$ such that

(i) $f(0) = 0$, $f'(u)$ is bounded in $0 \leqq u < \infty$;

(ii) $\int_0^\infty f(u)\, du$ exists and is equal to 1,

as may be seen. In particular, conditions (i) and (ii) are fulfilled by the first nonlinearity in (1.6) if either $p = 1$ or 2, or $p \geqq 3$ and $q > p + 1$, and by the second if $m = 0$, $a > 0$, and $p = 1$ or 2 or $p \geqq 3$, after multiplication by an appropriate positive constant.

5. Conclusions. In § 2 we showed that problem (1.1)-(1.5) has a unique classical solution in $0 \leqq t < T_0$, with $T_0 = \infty$ if $m \leqq 0$. If $0 < m < 1$, then $T_0 = \infty$ for appropriate initial conditions, but the solution is not expected to exist in $0 \leqq t < \infty$ for arbitrary initial data, as pointed out in Remark 2.2A.

Global stability properties for $m \neq 0$ were considered in § 3, where some previous partial numerical results on linear stability were confirmed and extended. In particular, the unique spatially one-dimensional steady state of (1.1)-(1.4) was shown to be unstable if $0 < m < \frac{1}{2}$ and globally, asymptotically stable in a certain sense if $m < 0$; in the latter case, it was shown also that (1.1)-(1.4) does not have other steady states, depending on the x_2 and/or the x_3 coordinates.

In § 3 we obtained sufficient conditions on the initial data for the solution of (1.1)-(1.5) to approach a given one-dimensional steady state.

Finally, let us point out that some questions about existence of more steady states and about the dynamics of (1.1)-(1.5) for $m \geqq 0$ remain unsolved. It seems that their solution requires more powerful mathematical tools (and perhaps some numerics on the two- and three-dimensional problems to get predictions) than those used in this paper. We think that any effort towards a complete understanding of (1.1)-(1.5) is worthwhile since, as was pointed out in the Introduction, Liñán's problem is ubiquitous in combustion theory.

Appendix. Spatially one-dimensional steady states of (1.1)-(1.4). We consider the one-dimensional steady states of (1.1)-(1.4) that satisfy the (slightly more general if $m \leqq 0$) boundary value problem

$$u'' = (u/2)\exp(mx - u) \quad \text{in } -\infty < x < \infty, \tag{A.1}$$

$$u \text{ bounded at } x = -\infty, \qquad |u - x| \text{ bounded at } x = +\infty, \tag{A.2}$$

where $u > 0$ in $\mathbb{R}$.

If $m \neq 0$, for each constant θ such that $0 < \theta < \infty$, (A.1) is invariant under the transformation

$$x \to \theta x - (2/m)\ln\theta, \qquad m \to m/\theta, \tag{A.3}$$

while the boundary conditions (A.2) become

$$u \text{ bounded at } x = -\infty, \qquad |u - \theta x| \text{ bounded at } x = \infty. \tag{A.4}$$

Therefore, problem (A.1)-(A.4), which will be considered below for convenience, is not essentially more general than (A.1), (A.2).

LEMMA A.1. *Every positive solution of* (A.1), (A.4) *satisfies*

$$\begin{aligned} &u \to u_0, \quad u' \to 0 \quad \text{as } x \to -\infty, \quad 0 < u' < \theta \quad \text{in } -\infty < x < \infty, \\ &u - \theta x \to c, \qquad u' \to \theta \quad \text{as } x \to +\infty \end{aligned} \tag{A.5}$$

for some finite constants $u_0 \geqq 0$ *and* c, *with* $u_0 = 0$ *if* $m \leqq 0$.

Proof. Since $u'' > 0$ in $-\infty < x < \infty$, the function $x \to u'(x)$ is strictly increasing, and the limits of u' at $x = -\infty$ and $x = +\infty$ exist; these limits are zero and θ, respectively, for (A.4) to be satisfied. Then $u' > 0$ in $-\infty < x < \infty$, and the limit of u at $x = -\infty$ exists, and it vanishes if $m \leqq 0$, for (A.4) to be satisfied. Finally, since $0 < u' < \theta$ in $-\infty < x < \infty$, the function $x \to u(x) - \theta x$ is strictly decreasing, and bounded at $x = +\infty$, and thus it must have a finite limit.

We first consider the case $m = 0$.

THEOREM A.2. *If* $m = 0$ *and* $\theta = 1$, *then* (A.1), (A.4) *possess a solution* u *that is unique up to translations, and such that*

$$u' > 0, \qquad u'^2 = 1 - (1+u)\exp(-u) \quad \text{in } -\infty < x < \infty. \tag{A.6}$$

If $m = 0$ *and* $0 < \theta < \infty$, $\theta \neq 1$, *then* (A.1)-(A.4) *has no solution.*

Proof. Equation (A.6) is obtained after multiplication of (A.1) by u' and integration from $-\infty$ to x, when taking into account (A.5). A further integration of (A.6) easily yields the desired result by phase-plane arguments.

The case $m<0$ is considered next. We first prove the following uniqueness result, which is used in the proof of Theorem 3.1 in § 3.

LEMMA A.3. *If $m<0$ and $0<\theta<\infty$, then there are not two distinct solutions of* (A.1), (A.5), u_1 *and* u_2, *such that* $u_1(x)\leqq u_2(x)$ *in* $-\infty<x<\infty$.

Proof. Suppose, on the contrary, that $u_1\not\equiv u_2$, and define $U_i(x)=u_i'(x)^2/2-F(u_i(x))\exp(mx)$, for $i=1$ and 2, where $F(u)=[1-(1+u)\exp(-u)]/2$. Then $U_i'(x)=-mF(u_i(x))\exp(mx)$, and

$$U_i(x)=-m\int_{-\infty}^{x}F(u_i(x))\exp(mx)\,dx$$

(the improper integral is seen to exist). Since the function F is strictly increasing, the function $x\to U_2(x)-U_1(x)$, which does not vanish identically, is nonnegative and increasing. Therefore, $\lim U_1(x)<\lim U_2(x)$ as $x\to\infty$, and this is not possible since $\lim U_1(x)=\lim U_2(x)=\theta^2/2$ as $x\to\infty$, from condition (A.5).

THEOREM A.4. *If $m<0$ and $0<\theta<\infty$, then* (A.1), (A.4) *have a unique solution u such that $u'(x)>0$ in $-\infty<x<\infty$, $u(x)\to 0$ as $x\to-\infty$, $u(x)-\theta x\to c$ as $x\to\infty$, for some finite constant c.*

Proof. If $\theta=1$, the result is readily obtained from Corollary 3.2 and Lemma A.1; if $\theta\neq 1$, the result is obtained by means of the transformation (A.3).

Remark A.5. Theorem A.4 contains the results by Hastings and Poore [18], who proved existence and uniqueness of solution of (A.1) for $m<0$, with boundary conditions

$$u'(x)\to 0 \quad \text{as } x\to-\infty, \qquad u'(x)\to\theta \quad \text{as } x\to\infty, \tag{A.7}$$

since, as we will see now, conditions (A.4) and (A.7) are equivalent when applied to (A.1). In fact, we will see that both (A.4) and (A.7) are equivalent to the following boundary condition:

$$u(x)\to 0 \quad \text{as } x\to-\infty, \qquad u(x)-\theta x\to c \quad \text{as } x\to\infty \tag{A.8}$$

for some finite constant c. That (A.4) implies (A.7) and (A.8) comes from Lemma A.1. Formulae (A.7) imply (A.4) since, by the argument in the proof of Lemma A.1, any solution u of (A.1), (A.7) satisfies (A.5). Furthermore, the function $x\to u(x)-\theta x$ is decreasing and thus it is bounded above as $x\to\infty$. Then, $u(x)\leqq w_2(x)$ in $-\infty<x<\infty$, where w_2 is the supersolution of (3.2), if b is sufficiently large. By the argument of the proof of Theorem 3.1, $u(x)\leqq\tilde{u}(x)$ in $-\infty<x<\infty$, where $\tilde{u}$ is the unique solution of (A.1), (A.4). Since $\tilde{u}$ satisfies (A.5), $u=\tilde{u}$ (Lemma A.3) and satisfies (A.4). Finally, any solution of (A.1), (A.8) clearly satisfies (A.4).

Now we consider the case $m>0$.

LEMMA A.6. *If $m>0$, for each $u_0\geqq 0$, there is a unique solution, $u(u_0;x)$, defined in $-\infty<x<\infty$, of the initial value problem*

$$\partial^2u/\partial x^2=(u/2)\exp(mx-u), \quad u\to u_0, \quad \partial u/\partial x\to 0 \quad \text{as } x\to-\infty, \tag{A.9}$$

and it is such that

(a) $u(0;x)=0$ *for all* $x\in\mathbb{R}$; $\partial u/\partial x>0$ *for all* $u_0>0$ *and all* $x\in\mathbb{R}$.

(b) *If* $u_0>0$, *then the derivative* $\partial u(u_0;x)/\partial u_0=z(u_0;x)$ *exists in* $-\infty<x<\infty$ *and is twice continuously differentiable with respect to* x.

(c) *If* $u_0 > 0$, *then the limits* $\lim \partial u(u_0; x)/\partial x = \psi(u_0)$ *and* $\lim \partial z(u_0; x)/\partial x = h(u_0)$, *as* $x \to \infty$, *exist. In addition, the function* $u_0 \to \psi(u_0)$ *is continuously differentiable and satisfies* $\psi(u_0) > m$, $\psi'(u_0) = h(u_0) \neq 0$ *in* $0 < u_0 < \infty$.

Proof. See Hastings and Poore [19], where this result is used to obtain uniqueness for (A.1), (A.7) when $m > 0$ and $\theta = 1$.

LEMMA A.7. *If* $m > 0$, *the function* ψ *of Lemma* A.6 *is such that*

(a) $\psi(u_0) \to 2m$ *as* $u_0 \to \infty$;

(b) $\psi(u_0) \to \infty$ *as* $u_0 \to 0$;

(c) $\psi'(u_0) < 0$ *for all* $u_0 > 0$.

Proof. (a) We multiply (A.9) by $\partial u/\partial x$, integrate from zero to $+\infty$, and integrate by parts twice, to obtain (recall that $\partial u(u_0; \infty)/\partial x > m$),

$$\begin{aligned}[\psi(u_0) + \partial u(u_0; 0)/\partial x - 2m][\psi(u_0) - \partial u(u_0; 0)/\partial x] \\ = [1 + u(u_0; 0)] \exp[-u(u_0; 0)] \\ + m \int_0^\infty \exp[mx - u(u_0; x)]\, dx.\end{aligned} \tag{A.10}$$

(Equation (A.10) was used by Ludford, Yanitell, and Buckmaster [5] to prove that (A.1), (A.7) has no solution if $\theta = 1$ and $\frac{1}{2} \leqq m < 1$.)

Now, since the function $x \to u(u_0; x)$ is strictly increasing, we have

$$\begin{aligned}\int_0^\infty \exp(mx - u)\, dx &< \int_0^\infty (u/u_0) \exp(mx - u)\, dx \\ &= (2/u_0)[\psi(u_0) - \partial u(u_0; 0)/\partial x]\end{aligned} \tag{A.11}$$

(use (A.9) to obtain the last equality). In addition, we multiply (A.9) by $\partial u/\partial x$, and integrate from $-\infty$ to zero, to obtain

$$\begin{aligned}[\partial u(u_0; 0)/\partial x]^2 &= \int_{-\infty}^0 u(\partial u/\partial x) \exp(mx - u)\, dx \\ &< \int_{-\infty}^0 u(\partial u/\partial x) \exp(-u)\, dx \\ &< \int_{u_0}^\infty u \exp(-u)\, du.\end{aligned}$$

Then

$$\partial u(u_0; 0)/\partial x \to 0 \quad \text{as } u_0 \to \infty, \tag{A.12}$$

and the desired result is easily obtained from (A.12), when we take into account (A.11), (A.12).

(b) For each $u_0 < 1$, let $x_1 \in \mathbb{R}$ be (uniquely) defined by $u(u_0; x_1) = 1$. Then $u(u_0; x) < 1$ for $x < x_1$ and integration of (A.9) from $-\infty$ to x yields

$$\begin{aligned}\partial u(u_0; x)/\partial x &= \int_{-\infty}^x (u/2) \exp(mx - u)\, dx \\ &< (u/2) \exp(-u) \int_{-\infty}^x \exp(mx)\, dx \\ &= (u/2m) \exp(mx - u)\end{aligned}$$

for all $x \in]-\infty, x_1[$ (the function $u \to u \exp(-u)$ is strictly increasing in $0 \leqq u \leqq 1$). Then $(2m/u)(\partial u/\partial x) \exp(u) < \exp(mx)$ in $-\infty < x < x_1$, and integration of this inequality from $-\infty$ to x_1 leads to

$$2m^2 \int_{u_0}^{1} u^{-1} \exp(u)\, du < \exp(mx_1).$$

Thus $x_1 \to \infty$ as $u_0 \to 0$, and the conclusion follows from the next equation, which is obtained by multiplying (A.9) by $\partial u/\partial x$ and integrating from x_1 to ∞

$$\psi(u_0)^2 - [\partial u(u_0; x_1)/\partial x]^2 = \int_{x_1}^{\infty} u(\partial u/\partial x) \exp(mx - u)\, dx$$
$$> \exp(mx_1) \int_{1}^{\infty} u \exp(-u)\, du.$$

(c) Since $\psi'(u_0) \neq 0$ in $0 < u_0 < \infty$ (Lemma A.6), parts (a) and (b) above yield the result.

THEOREM A.8. (a) *If $m > 0$ and $2m < \theta < \infty$, then* (A.1), (A.4) *have a unique solution u, and $u'(x) > 0$ in $-\infty < x < \infty$, $u(x) \to u_0$ as $x \to -\infty$ and $u(x) - \theta x \to c$ as $x \to \infty$, for some finite constants u_0 and c such that $u_0 > 0$. If $0 < \theta \leqq 2m$, then* (A.1), (A.4) *have no solution.*

(b) *If $m > 0$, let u_1 and u_2 be the solutions of* (A.1), (A.4) *for $\theta = \theta_1$ and $\theta = \theta_2$, with $2m < \theta_1 < \theta_2 < \infty$. Then $u_2 < u_1$ as $x \to -\infty$.*

Proof. Apply Lemmas A.1 and A.7.

Remarks A.9. (a) Part (b) of Theorem A.8 is needed in § 3 to analyze the asymptotic behavior of some solutions of (1.1)-(1.4) as $t \to \infty$ when $m > 0$.

(b) Part (a) of Theorem A.8 contains the results by Hastings and Poore [19], who proved existence and uniqueness of the solution of (A.1), (A.7) for $\theta = 1$ if $0 < m < \frac{1}{2}$ and nonexistence if $m \geqq \frac{1}{2}$, since, as in Remark A.5, conditions (A.4), (A.7), and

$$u \to u_0 \quad \text{as } x \to -\infty, \qquad u - \theta x \to c \quad \text{as } x \to \infty, \tag{A.13}$$

are equivalent when applied to (A.1). The equivalence of conditions (A.4), (A.7), and (A.13) can be proved by the same argument in Remark A.5 using Lemmas A.6 and A.7 and Theorem A.8.

The following result is needed in the proof of Theorem 3.4.

LEMMA A.10. *If $0 < m < \frac{1}{2}$, let u be the unique solution of* (A.1), (A.2). *If a solution $\bar{u}$ of* (A.1) *is such that*

$$0 \leqq \bar{u}(x) < u(x) \quad \textit{for all } x \in \mathbb{R},$$

then $\bar{u}(x) = 0$ for all $x \in \mathbb{R}$.

Proof. Since the function $x \to \bar{u}'(x)$ is strictly increasing, there exist the limits of $\bar{u}'$ as $x \to -\infty$ and as $x \to \infty$, and $\lim \bar{u}'(x) = 0$ as $x \to -\infty$, $\lim \bar{u}'(x) = \bar{\theta}$ as $x \to \infty$, for some $\bar{\theta} \in [0, 1]$. Then $\bar{u}$ satisfies (A.1), (A.4) (Remark A.9b), the limit of $\bar{u}$ as $x \to -\infty$, $\bar{u}_0$, exists and is finite (Lemma A.1), and $\bar{\theta}$ cannot be strictly positive (Theorem A.8). Therefore, $\bar{u}_0 = 0$ (Lemma A.6c) and $\bar{u}(x) = 0$ for all $x \in \mathbb{R}$ (Lemma A.6a).

REFERENCES

[1] E. CONWAY, D. HOFF, AND J. SMOLLER, *Large time behavior of solutions of systems of nonlinear reaction-diffusion equations*, SIAM J. Appl. Math., 35 (1978), pp. 1-16.

[2] A. LIÑÁN, *The asymptotic structure of counterflow diffusion flames for large activation energy*, Acta Astronaut., 1 (1974), pp. 1007-1039.

[3] ———, *Monopropellant droplet decomposition for large activation energies*, Acta Astronaut., 2 (1975), pp. 1009-1029.

[4] G. S. S. LUDFORD, D. W. YANNITELL, AND J. D. BUCKMASTER, *The decomposition of a hot monopropellant in an inert atmosphere*, Combust. Sci. Tech., 14 (1976), pp. 125-131.
[5] ———, *The decomposition of a cold monopropellant in an inert atmosphere*, Combust. Sci. Tech., 14 (1976), pp. 133-145.
[6] M. MATALON, G. S. S. LUDFORD, AND J. D. BUCKMASTER, *Diffusion flames in a chamber*, Acta Astronaut., 6 (1979), pp. 943-959.
[7] A. K. KAPILA AND G. S. S. LUDFORD, *Two step sequential reaction for large activation energy*, Combust. Flame, 29 (1977), pp. 167-176.
[8] J. PELÁEZ, *Stability of premixed flames with two thin reaction layers*, SIAM J. Appl. Math., 47 (1987), pp. 781-799.
[9] A. K. KAPILA AND A. B. POORE, *The steady response of a nonadiabatic tubular reactor: new multiplicities*, Chem. Engrg. Sci., 37 (1982), pp. 57-68.
[10] F. A. WILLIAMS, *Combustion Theory*, Benjamin-Cummings, Menlo Park, CA, 1985.
[11] G. I. SIVASHINSKY, *Nonlinear analysis of hydrodynamic instability in laminar flames.* I: *Derivation of basic equations*, Acta Astronaut., 4 (1977), pp. 1177-1206.
[12] B. J. MATKOWSKY AND G. I. SIVASHINSKY, *An asymptotic derivation of two models in flame theory associated with the constant density approximation*, SIAM J. Appl. Math., 37 (1979), pp. 686-699.
[13] G. JOULIN AND P. CLAVIN, *Linear stability analysis of non-adiabatic flames: Diffusional-thermal model*, Combust. Flame, 35 (1979), pp. 139-153.
[14] A. K. KAPILA AND B. J. MATKOWSKY, *Reactive-diffusive systems with Arrhenius kinetics: Multiple solutions, ignition and extinction*, SIAM J. Appl. Math., 36 (1979), pp. 373-389.
[15] A. K. KAPILA, B. J. MATKOWSKY, AND J. M. VEGA, *Reactive-diffusive systems with Arrhenius kinetics: Peculiarities of the spherical geometry*, SIAM J. Appl. Math., 38 (1980), pp. 382-341.
[16] J. M. VEGA AND A. LIÑÁN, *Singular Langmuir-Hinshelwood reaction-diffusion problems: Strong adsorption under quasi-isothermal conditions*, SIAM J. Appl. Math., 42 (1982), pp. 1047-1068.
[17] J. M. VEGA, *Singular Langmuir-Hinshelwood reaction-diffusion problems: Strongly nonisothermal conditions*, SIAM J. Appl. Math., 43 (1983), pp. 1367-1389.
[18] S. P. HASTINGS AND A. B. POORE, *A nonlinear problem arising from combustion theory: Liñán's problem*, SIAM J. Math. Anal., 14 (1983), pp. 425-430.
[19] ———, *Liñán's problem from combustion theory, part* II, SIAM J. Math. Anal., 16 (1985), pp. 331-340.
[20] N. PETERS, *On the stability of Liñán's premixed flame regime*, Combust. Flame, 33 (1978), pp. 315-318.
[21] D. S. STEWART AND J. D. BUCKMASTER, *The stability of Liñán's premixed flame regime revisited*, SIAM J. Appl. Math., 46 (1986), pp. 582-587.
[22] J. D. BUCKMASTER, A. NACHMAN, AND S. D. TALIAFERRO, *The fast time instability of diffusion flames*, Phys. D, 9 (1983), pp. 408-424.
[23] S. D. TALIAFERRO, *Stability and bifurcation of equilibrium solutions of reaction-diffusion equations on an infinite space interval*, J. Differential Equations, 54 (1984), pp. 19-59.
[24] O. A. LADYZENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.
[25] S. D. EIDEL'MAN, *Parabolic Systems*, North-Holland, Amsterdam, New York, 1969.
[26] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.
[27] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[28] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[29] C. ALVAREZ PEREIRA AND J. M. VEGA, *On the large-time behavior of a premixed, non-adiabatic reaction zone*, in preparation.
[30] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, New York, 1983.
[31] M. W. HIRSCH, *The dynamical systems approach to differential equations*, Bull. Amer. Math. Soc. (N.S.), 11 (1984), pp. 1-64.
[32] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
[33] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, New York, 1981.
[34] P. C. FIFE, *Asymptotic states for equations of reaction and diffusion*, Bull. Amer. Math. Soc., 84 (1978), pp. 693-726.
[35] M. W. HIRSCH, *Stability and convergence in strongly monotone flows*, preprint, Center for Pure and Applied Mathematics, University of California, Berkeley, CA, 1984.
[36] N. D. ALIKAKOS AND P. BATES, *Stabilization of solutions for a class of degenerate equations in divergence form in one space dimension*, preprint, 1988.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 905-916, July 1990
© 1990 Society for Industrial and Applied Mathematics
005

A SYSTEM OF DEGENERATE PARABOLIC EQUATIONS*

M. BERTSCH† AND S. KAMIN‡

Abstract. The system of two nonlinear equations which arises in plasma physics is considered. The equations are of degenerate parabolic type. The global existence theorem for the Cauchy problem is proved. The proof is based on the Lagrangian transformation, thus using a particular structure of the system.

Key words. system of nonlinear equations, degenerate equations, mass coordinate plasma physics

AMS(MOS) subject classifications. 35K65, 35K45

1. Introduction. Let $\rho(x, t)$ and $T(x, t)$ denote the density and ionic temperature of a plasma that slowly diffuses in a strong magnetic field. To study the effect of the nonlinear coupled diffusion of mass and heat in plasma, Rosenau and Hyman [RH1], [RH2] introduced the simplified, one-dimensional system of partial differential equations

$$\rho_t = (D_1(\rho, T)\rho_x)_x,$$
$$(\rho T)_t = (\rho D_2(\rho, T)T_x)_x + (TD_1(\rho, T)\rho_x)_x,$$

where D_1 and D_2 are power-type nonlinearities:

$$D_i(\rho, T) = d_i\rho^{\alpha_i}T^{\beta_i} \quad \text{for } \rho, T \geqq 0, \quad i = 1, 2. \tag{1.1}$$

Here d_i, α_i, and β_i are constants that depend on the properties of the particular physical process. In the isothermal case, i.e., where T is constant, the system reduces to a single equation for ρ, the so-called porous medium equation, which has been studied extensively in the literature (see, for example, [Ar]).

Observe that the equations are not uniformly parabolic: if, for example, $\alpha_i > 0$, D_i vanishes at $\rho = 0$. We say that the equations are of degenerate parabolic type. In this paper we are interested in a global existence result. However, even for nondegenerate quasilinear systems, most of the results in the literature concern local existence [Am1], [Am2]. Recently, Amann [Am3] proved a global existence result for diagonal nondegenerate systems in divergence form. We also mention a paper by Alt and Luckhaus [AL] who consider a class of degenerate systems. However, this class does not contain the system studied here.

Below we define a weak solution of our degenerate system, and we prove the global existence of such a solution. To be more precise, we study the problem

$$\text{(I)}\begin{cases} \rho_t = (D_1\rho_x)_x & \text{in } Q \equiv \mathbb{R}\times\mathbb{R}^+, \quad (1.2)\\ (\rho T)_t = (\rho D_2 T_x)_x + (TD_1\rho_x)_x & \text{in } Q, \quad (1.3)\\ \rho(x, 0) = \rho_0(x) & \text{for } x \in \mathbb{R},\\ T(x, 0) = T_0(x) & \text{for } x \in \mathscr{P}_0 \subset \mathbb{R}, \end{cases}$$

where

$$\mathscr{P}_0 = \{x \in \mathscr{R}: \rho_0(x) > 0\}, \tag{1.4}$$

where ρ_0 and T_0 are nonnegative, bounded, and continuous functions on $\mathbb{R}$, respectively, $\mathscr{P}_0$, and where

$$T_0 \geqq \nu > 0 \quad \text{on } \mathscr{P}_0,$$

* Received by the editors February 22, 1989; accepted for publication August 2, 1989.

† Dipartimento di Matematica, Università di Torino, Torino, Italy.

‡ School of Mathematical Sciences, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel-Aviv, Israel.

for some constant ν. Instead of (1.1), we will assume that the functions $D_i = D_i(\rho, T)$ are of the form

$$D_i(\rho, T) = \rho^{\alpha_i}\varphi_i(\rho, T) \quad \text{for } \rho, T > 0, \tag{1.5}$$

where $a_1 > 0$, $a_2 > -1$ and where φ_1 and φ_2 are smooth functions on $[0, \infty] \times (0, \infty)$ which satisfy

$$\varphi_i > 0 \quad \text{in } [0, \infty) \times (0, \infty), \quad i = 1, 2. \tag{1.6}$$

It will turn out that the initial condition $T_0 \geqq \nu$ implies that the solution (ρ, T) of problem I which we will construct below satisfies $T \geqq \nu$ for all later times. Physically, the strict positivity of $T(x, t)$ seems to be consistent with the fact that the temperature in a plasma is high. Mathematically, however, combined with condition (1.6), it is a considerable restriction: it implies the strict positivity of $\varphi_i(\rho(x, t), T(x, t))$, which means that the degeneracy of the parabolicity in (1.2) and (1.3) is only caused by the ρ-dependence of $D_i(\rho, T)$.

In view of (1.3), a natural quantity to consider is the energy density $E = \rho T$. Since we assume that all physical quantities are bounded, it follows that E vanishes if $\rho = 0$. Therefore it seems reasonable not to prescribe T on the set where ρ vanishes. Physically, this means that we do not have to define the temperature in a vacuum. This will also be reflected in the definition of a solution: we will define the temperature T only in the positivity set

$$\mathcal{P} = \bigcup_{t \geqq 0} \mathcal{P}_t \tag{1.7}$$

where

$$\mathcal{P}_t = \{x \in \mathbb{R}: \rho(x, t) > 0\} \tag{1.8}$$

(for the definition of a solution we refer to § 2, where we also give the precise assumptions on the data). In [BK1] it has been proved that the set $\mathcal{P}_t$ does not shrink when time evolves.

Observe that mathematically the definition of the weak solution, and in particular our choice not to define T outside the set $\mathcal{P}$, can only be justified by proving the existence and uniqueness of such a solution. Unfortunately, the uniqueness problem is still completely open. It is clear, however, that uniqueness certainly should fail if we would define T outside $\bar{\mathcal{P}}$, because we could choose T in an arbitrary way outside $\bar{\mathcal{P}}$.

In §§ 3 and 4 we prove that problem I has a solution. In § 3 we reduce the problem to a nondegenerate problem in a bounded domain. The latter one we solve in § 4.

In the case of the power-type nonlinearities (1.1), Hyman and Rosenau have constructed separable solutions of a closely related system in a bounded domain, and they observe that the large-time behaviour of these solutions depends strongly on the parameters.

In a forthcoming paper [BK2] we will give a detailed analysis of the asymptotic behaviour of the solution (ρ, T) of problem I in the case where the functions φ_i in (1.5) do not depend on ρ. It turns out that this behaviour depends critically on the sign of $a_1 - a_2$, where a_1 and a_2 are defined by (1.5).

2. Preliminaries. Throughout the paper we will use the following assumptions on the data of problems I and II.

H_1. φ_1, $\varphi_2 \in C^2([0,\infty)\times(0,\infty))$ *and satisfy* $\varphi_i > 0$ *in* $[0,\infty)\times(0,\infty)$; $a_1 > 0$ *and* $a_2 > -1$; *the functions* D_1 *and* D_2 *are defined by* (1.5).

H_2. $\rho_0 \in C(\mathbb{R}) \cap L^\infty(\mathbb{R})$, $\rho_0 \geqq 0$ *and* $\rho_0 \not\equiv 0$ *on* $\mathbb{R}$, *and* $\mathscr{P}_0 \subset \mathbb{R}$ *is defined by* (1.4).

H_3. $T_0 \in C(\mathscr{P}_0) \cap L^\infty(\mathscr{P}_0)$ *and* $T_0 \geqq \nu$ *on* $\mathscr{P}_0$ *for some* $\nu > 0$.

Next we define what we mean by a solution of problems I and II.

DEFINITION 2.1. A pair (ρ, T) is called a solution of problem I if:

(i) $\rho \in C(\bar{Q}) \cap L^\infty(Q)$, $T \in C(\mathscr{P}) \cap L^\infty(\mathscr{P})$, where $\mathscr{P} \subset \bar{Q}$ is defined by (1.7);

(ii) $\rho \geqq 0$ in Q, $T \geqq \varepsilon$ in $\mathscr{P}$ for some $\varepsilon > 0$;

(iii) $(\rho^{a_1+1/2})_x \in L^2_{\mathrm{loc}}(\bar{Q})$, $\rho^{(a_2+1)/2} T_x \in L^2(S)$ for any bounded set $S \subset \mathscr{P}$;

(iv) for any $\psi \in C^{1,1}(\bar{Q})$ with compact support

$$\int_{\mathscr{P}_0} \rho_0(x)\psi(x,0)\,dx + \iint_{\mathscr{P}} \{\rho\psi_t - \rho^{a_1}\varphi_1(\rho,T)\rho_x\psi_x\}\,dx\,dt = 0, \tag{2.1}$$

$$\int_{\mathscr{P}_0} \rho_0(x)T_0(x)\psi(x,0)\,dx + \iint_{\mathscr{P}} \{\rho T\psi_t - \rho^{a_2+1}\varphi_2(\rho,T)T_x\psi_x - \rho^{a_1}T\varphi_1(\rho,T)\rho_x\psi_x\}\,dx\,dt = 0, \tag{2.2}$$

where $\mathscr{P}_0$ is defined by (1.4).

3. Existence. In the present and following section we prove that problem I has a solution.

THEOREM 3.1. *Let hypotheses* H_1-H_3 *be satisfied. Then problem* I *possesses a solution* (ρ, T). *In addition,* ρ *and* T *are classical solutions in the set* $\mathring{\mathscr{P}}$, *i.e.,* ρ, $T \in C^{2,1}(\mathring{\mathscr{P}})$ *and satisfy* (1.2) *and* (1.3) *in* $\mathring{\mathscr{P}}$.

Remark. In [BK1] it has been proved that the solution which we construct in Theorem 3.1 has the property that the set $\mathscr{P}_t = \mathscr{P} \cap \{t\}$ is not shrinking in time. In particular, if $\rho_0 > 0$ in $\mathbb{R}$, then $\mathscr{P} = \mathbb{R}\times[0,\infty)$, and thus, by Theorem 3.1, ρ and T are classical solutions.

To prove Theorem 3.1 we first approximate problem I by a system of nondegenerate parabolic equations. In § 4 we will prove existence for this nondegenerate system.

We introduce some notation:

$$\rho^* = \sup\{\rho_0(x), x \in \mathbb{R}\}, \qquad T^* = \sup\{T_0(x), x \in \mathscr{P}_0\}. \tag{3.1}$$

Proof of Theorem 3.1. We fix an arbitrary $\tau > 0$. It is enough to prove existence for $t \in [0,\tau]$.

For any $\varepsilon \in (0, \frac{1}{2}\rho^*)$ there exists a constant $L_\varepsilon > 0$ and functions $\rho_{0\varepsilon}$, $T_{0\varepsilon} \in C^\infty([-L_\varepsilon, L_\varepsilon])$ such that

(i) $\varepsilon \leqq \rho_{0\varepsilon} \leqq \rho^*$ and $\nu \leqq T_{0\varepsilon} \leqq T^*$ in $[-L_\varepsilon, L_\varepsilon]$,

(ii) $L_\varepsilon \to \infty$ as $\varepsilon \searrow 0$;

(iii) $\rho_{0\varepsilon} \to \rho_0$ in $C_{\mathrm{loc}}(\mathbb{R})$ as $\varepsilon \searrow 0$ and $T_{0\varepsilon} \to T_0$ in $C_{\mathrm{loc}}(\mathscr{P}_0)$ as $\varepsilon \searrow 0$;

(iv) $\rho'_{0\varepsilon}(\pm L_\varepsilon) = T'_{0\varepsilon}(\pm L_\varepsilon) = 0$.

We consider the approximating problem:

$$(\mathrm{I}_\varepsilon)\begin{cases} \rho_t = (D_1\rho_x)_x & \text{in } Q_{\varepsilon,\tau} \equiv (-L_\varepsilon, L_\varepsilon)\times(0,\tau] \\ (\rho T)_t = (\rho D_2 T_x)_x + (T D_1 \rho_x) & \text{in } Q_{\varepsilon,\tau} \\ \rho_x(\pm L_\varepsilon, t) = T_x(\pm L_\varepsilon, t) = 0 & \text{for } 0 < t \leqq \tau \\ \rho(x,0) = \rho_{0\varepsilon}(x) & \text{for } -L_\varepsilon < x < L_\varepsilon \\ T(x,0) = T_{0\varepsilon}(x) & \text{for } -L_\varepsilon < x < L_\varepsilon. \end{cases}$$

Since $\rho_{0\varepsilon} \geqq \varepsilon > 0$, the maximum principle implies that a smooth solution of problem I_ε will satisfy $\rho \geqq \varepsilon$ in Q_ε. Therefore, problem I_ε is no longer degenerate parabolic. This will be used to prove the following existence result for problem I_ε.

LEMMA 3.2. *Let $\rho_{0\varepsilon}$ and $T_{0\varepsilon}$ be as above. Then there exists a classical solution $(\rho_\varepsilon, T_\varepsilon)$ of problem I_ε, i.e., there exist functions ρ_ε, $T_\varepsilon \in C^{2,1}(\bar{Q}_{\varepsilon,\tau})$ which satisfy pointwise the equations and boundary and initial conditions of problem I_ε.*

The proof will be given in § 4.

We establish some a priori estimates for $(\rho_\varepsilon, T_\varepsilon)$.

LEMMA 3.3. *Let ρ_ε and t_ε be defined by Lemma 3.2. Then*

(i) $\varepsilon \leqq \rho_\varepsilon \leqq \rho^*$ *and* $\nu \leqq T_\varepsilon \leqq T^*$ *in* $Q_{\varepsilon,\tau}$;

(ii) *There exists a constant $\mathscr{C} > 0$ that does not depend on ε such that for any $0 < L < L_\varepsilon - 1$*

$$\iint_{(-L,L)\times(0,\tau)} (\rho_\varepsilon^{a_1+1/2})_x^2 \, dx\, dt \leqq \mathscr{C}\left\{\int_{-L-1}^{L+1} \rho_{0\varepsilon}^{a_1+1} \, dx + 1\right\}, \tag{3.2}$$

$$\iint_{(-L,L)\times(0,\tau)} \rho_\varepsilon^{a_2+1} T_{\varepsilon x}^2 \, dx\, dt \leqq \mathscr{C}\left\{\int_{-L-1}^{L+1} \rho_{0\varepsilon} \, dx + 1\right\}. \tag{3.3}$$

Proof. Observe that the two equations in the system can be written as

$$\rho_t = (\rho^{a_1}\varphi_1(\rho, T)\rho_x)_x, \tag{3.4}$$

$$\rho T_t = (\rho^{a_2+1}\varphi_2(\rho, T)T_x)_x + \rho^{a_1}\varphi_1(\rho, T)\rho_x T_x. \tag{3.5}$$

Hence part (i) follows at once from the maximum principle.

Let $\psi \in C^2(\mathbb{R})$ satisfy

$$\begin{aligned} &0 \leqq \psi \leqq 1 \quad \text{on } \mathbb{R}; \qquad |\psi_x| \leqq 2 \quad \text{on } \mathbb{R}; \\ &\psi(x) = 0 \quad \text{if } |x| \geqq L+1; \qquad \psi(x) = 1 \quad \text{if } |x| \leqq L. \end{aligned} \tag{3.6}$$

We multiply equation (3.4) by $\rho_\varepsilon^{a_1}\psi^2$ and integrate by parts over $Q_{\varepsilon,\tau}$. Then

$$\begin{aligned} &\frac{a_1}{(a_1+\frac{1}{2})^2}\iint_{Q_{\varepsilon,\tau}} \varphi_1(\rho_\varepsilon^{a_1+1/2})_x^2\psi^2 \, dx\, dt + \frac{1}{a_1+1}\int_{-L_\varepsilon}^{L_\varepsilon} \psi^2(x)\rho_\varepsilon^{a_1+1}(x,\tau)\, dx \\ &= \frac{1}{a_1+1}\int_{-L_\varepsilon}^{L_\varepsilon} \psi^2\rho_{0\varepsilon}^{a_1+1} \, dx - \frac{2}{a_1+\frac{1}{2}}\iint_{Q_{\varepsilon,\tau}} \varphi_1(\rho_\varepsilon^{a_1+1/2})_x\rho_\varepsilon^{a_1+1/2}\psi\psi_x \, dx\, dt, \end{aligned} \tag{3.7}$$

and, if we apply Cauchy-Schwarz and Young's inequality to the last term at the right-hand side, and use the properties of ψ given by (3.6), it follows easily that

$$\iint_{Q_{\varepsilon,\tau}} (\rho_\varepsilon^{a_1+1/2})_x^2\psi^2 \, dx\, dt \leqq \mathscr{C}\left\{\int_{-L_\varepsilon}^{L_\varepsilon} \psi^2\rho_{0\varepsilon}^{a_1+1} \, dx + 1\right\}. \tag{3.8}$$

Clearly this implies (3.2).

Finally, to prove (3.3), we multiply (3.5) by $T_\varepsilon\psi^2$. Integrating by parts over $Q_{\varepsilon,\tau}$ we find that

$$\begin{aligned} &\iint_{Q_{\varepsilon,\tau}} \psi^2\rho_\varepsilon^{a_2+1}\varphi_2 T_{\varepsilon x}^2 \, dx\, dt + \frac{1}{2}\int_{-L_\varepsilon}^{L_\varepsilon} \psi^2(x)\rho_\varepsilon(x,\tau)T_\varepsilon^2(x,\tau)\, dx \\ &= \frac{1}{2}\int_{-L_\varepsilon}^{L_\varepsilon} \psi^2\rho_{0\varepsilon}T_{0\varepsilon}^2 \, dx + \iint_{Q_{\varepsilon,\tau}} \psi^2\left\{\frac{1}{2}\rho_{\varepsilon t}T_\varepsilon^2 + \rho_\varepsilon^{a_1}\varphi_1\rho_{\varepsilon x}T_\varepsilon T_{\varepsilon x}\right\} dx\, dt \\ &\quad -2\iint_{Q_{\varepsilon,\tau}} \rho_\varepsilon^{a_2+1}\varphi_2 T_\varepsilon T_{\varepsilon x}\psi\psi_x \, dx\, dt. \end{aligned} \tag{3.9}$$

Multiplying equation (3.4) by $\frac{1}{2}\psi^2T^2$ and integrating by parts over $Q_{\varepsilon,\tau}$, we find that the second term at the right-hand side of (3.9) is equal to

$$I \equiv \iint_{Q_{\varepsilon,\tau}} \rho_\varepsilon^{a_1}\varphi_1\rho_{\varepsilon x}T_\varepsilon^2\psi\psi_x \, dx\, dt,$$

which can be controlled by (3.8): since $a_1 > 0$ we obtain that $|I| \leqq \mathscr{C} + \mathscr{C} \int_{-L_\varepsilon}^{L_\varepsilon} \rho_{0\varepsilon} \psi^2 \, dx$. Again applying Cauchy-Schwarz and Young's inequality to the last term in (3.9), and using that $a_2 > -1$, we arrive at (3.3), and the proof of Lemma 3.3 is complete.

Using Lemmas 3.2 and 3.3 we can construct a solution (ρ, T) of problem I.

Let $(\rho_\varepsilon, T_\varepsilon)$ be defined by Lemma 3.2. By Lemma 3.3(i), $\{\rho_\varepsilon\}$ is a family of equibounded solutions of the equation

$$\rho_t = (A_\varepsilon(x, t)\rho^{a_1}\rho_x)_x$$

where

$$A_\varepsilon(x, t) \equiv \varphi_1(\rho_\varepsilon(x, t), T_\varepsilon(x, t)).$$

Since

$$0 < A^- \leqq A_\varepsilon(x, t) \leqq A^+$$

for some constants A^- and A^+ independent of ε, it follows from a result of DiBenedetto [dB] that the functions ρ_ε are equicontinuous on bounded subsets of $\mathbb{R} \times [0, \tau]$. Hence there exists a sequence $\varepsilon_i \searrow 0$ as $i \to \infty$ and a nonnegative, bounded, and continuous function ρ on $\mathbb{R} \times [0, \tau]$, such that

$$\rho_{\varepsilon_i} \to \rho \text{ uniformly on compact subsets of } \mathbb{R} \times [0, \tau] \quad \text{as } i \to \infty. \tag{3.10}$$

Let $Q_\tau = \mathbb{R} \times (0, \tau]$, and let $\mathscr{P}$ be defined by (1.7). In view of the estimates in Lemma 3.3 we may assume that the sequence ε_i is chosen such that for all $p \in [1, \infty)$

$$T_{\varepsilon_i} \to T \text{ weakly in } L^p_{\text{loc}}(\bar{Q}_\tau) \quad \text{as } i \to \infty \tag{3.11}$$

for some bounded and nonnegative function T on Q_τ, and that

$$(\rho^{a_1+1/2})_x \in L^2_{\text{loc}}(\bar{Q}_\tau) \quad \text{and} \quad (\rho_{\varepsilon_i}^{a_1+1/2})_x \to (\rho^{a_1+1/2})_x \text{ weakly in } L^2_{\text{loc}}(\bar{Q}_\tau) \quad \text{as } i \to \infty, \tag{3.12}$$

and

$$T_x \in L^2_{\text{loc}}(\mathscr{P} \cap \bar{Q}_\tau) \quad \text{and} \quad T_{\varepsilon_i x} \to T_x \text{ weakly in } L^2_{\text{loc}}(\mathscr{P} \cap \bar{Q}_\tau) \quad \text{as } i \to \infty. \tag{3.13}$$

We claim that the pair (ρ, T) is a solution of problem I on $[0, \tau]$.

LEMMA 3.4. *The functions T_ε, defined by Lemma 3.2, are locally uniformly continuous in $\mathscr{P} \cap \bar{Q}_\tau$.*

The proof will be given in § 4.

Remark. Locally in $\mathscr{P}$, ρ is bounded away from zero. Hence Lemma 3.4 is essentially a result about the nondegenerate system.

Lemma 3.4 implies that $T \in C(\mathscr{P} \cap \bar{Q}_\tau)$. Furthermore, we may assume that

$$T_{\varepsilon_i} \to T \quad \text{in } C_{\text{loc}}(\mathscr{P} \cap Q_\tau) \quad \text{as } i \to \infty. \tag{3.14}$$

It remains to prove that ρ and T satisfy the integral identities (2.1) and (2.2).

So let $\psi \in C^{1,1}(\bar{Q})$ have bounded support in $\mathbb{R} \times [0, \tau)$. Let ε_i be so small that $\operatorname{supp} \psi \subset \bar{Q}_{\varepsilon_i, \tau}$. In view of (3.10), (2.1) follows if we show that

$$\iint_{Q_{\varepsilon_i, \tau}} \rho_{\varepsilon_i}^{a_1} \varphi_1(\rho_{\varepsilon_i}, T_{\varepsilon_i}) \rho_{\varepsilon_i x} \psi_x \to \iint_{\mathscr{P} \cap [0,\tau]} \rho^{a_1} \varphi_1(\rho, T) \rho_x \psi_x \quad \text{as } i \to \infty. \tag{3.15}$$

Let $\delta > 0$ and define the sets U_δ^- and U_δ^+ by

$$\begin{aligned} U_\delta^- &= \{(x, t) \subset \operatorname{supp} \psi : 0 \leqq \rho(x, t) \leqq \delta\} \\ U_\delta^+ &= \{(x, t) \subset \operatorname{supp} \psi : \rho(x, t) > \delta\} \subset \mathscr{P}. \end{aligned} \tag{3.16}$$

By (3.10), there exists an $\varepsilon_0 > 0$ such that

$$0 < \rho_{\varepsilon_i} \leqq 2\delta \quad \text{in } U_\delta^- \cap Q_{\varepsilon_i,\tau} \quad \text{if } \varepsilon_i \leqq \varepsilon_0$$

and

$$\rho_{\varepsilon_i} \geqq \tfrac{1}{2}\delta \quad \text{in } U_\delta^+ \cap Q_{\varepsilon_i,\tau} \quad \text{if } \varepsilon_i \leqq \varepsilon_0.$$

Hence, by (3.2),

$$\left| \iint_{U_\delta^- \cap Q_{\varepsilon_i,\tau}} \rho_{\varepsilon_i}^{a_1} \varphi_1(\rho_{\varepsilon_i}, T_{\varepsilon_i}) \rho_{\varepsilon_i x} \psi_x \right| \leqq \mathscr{C}\sqrt{\delta} \quad \text{if } \varepsilon_i \leqq \varepsilon_0 \tag{3.17}$$

for some $\mathscr{C} > 0$ which may depend on supp ψ, but which does not depend on ε_i and δ.

In addition it follows from (3.10), (3.12), and (3.14) that, for $\delta > 0$ fixed,

$$\iint_{U_\delta^+ \cap Q_{\varepsilon_i,\tau}} \rho_{\varepsilon_i}^{a_1} \varphi_1(\rho_{\varepsilon_i}, T_{\varepsilon_i}) \rho_{\varepsilon_i x} \psi_x \to \iint_{U_\delta^+} \rho^{a_1} \varphi_1(\rho, T) \rho_x \psi_x \quad \text{as } i \to \infty,$$

and, combining this with (3.17) and letting $\delta \to 0$, we find (3.15).

From a similar procedure it follows that

$$\iint_{Q_{\varepsilon_i,\tau}} \{\rho_{\varepsilon_i} T_{\varepsilon_i} \psi_t - \rho_{\varepsilon_i}^{a_1} \varphi_1(\rho_{\varepsilon_i}, T_{\varepsilon_i}) T_{\varepsilon_i} \rho_{\varepsilon_i x} \psi_x\}$$
$$\to \iint_{\mathscr{P} \cap [0,\tau]} \{\rho T \psi_t - \rho^{a_1} \varphi_1(\rho, T) T \rho_x \psi_x\} \quad \text{as } i \to \infty.$$

Hence, to prove (2.2), we have to show that

$$\iint_{Q_{\varepsilon_i,\tau}} \rho_{\varepsilon_i}^{a_2+1} \varphi_2(\rho_{\varepsilon_i}, T_{\varepsilon_i}) T_{\varepsilon_i x} \psi_x \to \iint_{\mathscr{P} \cap [0,\tau]} \rho^{a_2+1} \varphi_2(\rho, T) T_x \psi_x \quad \text{as } i \to \infty. \tag{3.18}$$

Let $U_\delta^\pm$ be defined by (3.16). By (3.10), (3.13), and (3.14),

$$\iint_{U_\delta^+ \cap Q_{\varepsilon_i,\tau}} \rho_{\varepsilon_i}^{a_2+1} \varphi_2(\rho_{\varepsilon_i}, T_{\varepsilon_i}) T_{\varepsilon_i x} \psi_x \to \iint_{U_\delta^+} \rho^{a_2+1} \varphi_2(\rho, T) T_x \psi_x \quad \text{as } i \to \infty.$$

In addition there exist constants $\mathscr{C}_0, \mathscr{C}_1 > 0$ which do not depend on ε_i and δ, such that

$$\left| \iint_{U_\delta^- \cap Q_{\varepsilon_i,\tau}} \rho_{\varepsilon_i}^{a_2+1} \varphi_2(\rho_{\varepsilon_i}, T_{\varepsilon_i}) T_{\varepsilon_i x} \psi_x \right| \leqq \mathscr{C}_0 \left\{ \iint_{U_\delta^- \cap Q_{\varepsilon_i,\tau}} \rho_{\varepsilon_i}^{a_2+1} T_{\varepsilon_i x}^2 \right\}^{1/2} \left\{ \iint_{U_\delta^- \cap Q_{\varepsilon_i,\tau}} \rho_{\varepsilon_i}^{a_2+1} \right\}^{1/2}$$
$$\leqq \mathscr{C}_1 \delta^{(a_2+1)/2},$$

where we have used (3.3). Here $\mathscr{C}_1$ may depend on supp ψ. Since $a_2 > -1$ we arrive at (3.18) if we let $\delta \searrow 0$.

Finally we have to show that the solution (ρ, T) which we have constructed above is classical in $\mathring{\mathscr{P}}$, i.e., that $\rho, T \in C^{2,1}(\mathring{\mathscr{P}})$. This is an immediate consequence of the following lemma.

LEMMA 3.5. *The functions ρ_ε and T_ε, defined by Lemma* 3.2, *are equibounded in $C_{\text{loc}}^{2+\alpha,1+\alpha/2}(\mathring{\mathscr{P}})$ for $\alpha \in (0,1)$.*

Again the proof will be given in § 4.

4. The nondegenerate system. The main result of this section is that the nondegenerate system I_ε has a smooth solution, i.e., we will prove Lemma 3.2.

Most of the results in the literature about nonlinear, uniformly parabolic systems concern the local existence of solutions (by local we refer to local with respect to the time t). See for example [Am1]. For the global continuation of these solutions a priori bounds that are stronger than the available estimates are often needed.

To prove global existence for our system, we will exploit its particular structure. The main obstacle to prove global existence is the occurrence of the cross-diffusion term $(TD_1\rho_x)_x$ in the second equation. Physically this term represents the transport of heat due to the diffusion of mass. Therefore we expect that this term disappears if we use mass or Lagrangian coordinates. Indeed, defining the new coordinates $(y, \underline{t})$ by

$$\begin{aligned} y &= \int_{-L_\varepsilon}^{x} \rho(s, t)\, ds \\ \underline{t} &= t \end{aligned} \qquad \text{for } -L_\varepsilon \leqq x \leqq L_\varepsilon, \quad 0 \leqq t \leqq \tau, \tag{4.1}$$

we will see below that we end up with a system in diagonal form. Recently this Lagrangian transformation has been applied to obtain several results about scalar nonlinear diffusion equations. See, for example, [BvDEZ], [BdP].

Let $(y, \underline{t})$ be defined by (4.1). Then

$$0 \leqq y \leqq M_\varepsilon \equiv \int_{-L_\varepsilon}^{L_\varepsilon} \rho_{0\varepsilon}(x)\, dx \quad \text{for all } t \in [0, \tau], \tag{4.2}$$

where we have used the conservation of the "total mass" $\int_{-L_\varepsilon}^{L_\varepsilon} \rho_\varepsilon(s, t)\, ds$.

From now on we will omit the subscript ε. Assuming that ρ is a smooth solution of (1.2), we obtain from (4.1) that

$$y_x = \rho \quad \text{and} \quad y_t = D_1\rho_x = \rho D_1\rho_y. \tag{4.3}$$

Then

$$\rho_t = \rho_{\underline{t}} + \rho_y y_t = \rho_{\underline{t}} + \rho D_1\rho_y^2,$$

and, on the other hand

$$\rho_t = (D_1\rho_x)_x = (\rho D_1\rho_y)_y\rho = \rho^2(D_1\rho_y)_y + \rho D_1\rho_y^2.$$

Hence

$$\rho_{\underline{t}} = \rho^2(D_1\rho_y)_y \quad \text{in } G \equiv (0, M) \times (0, \tau], \tag{4.4}$$

or, since $\rho \geqq \varepsilon > 0$,

$$\left(\frac{1}{\rho}\right)_{\underline{t}} = \left(\rho^2 D_1\left(\frac{1}{\rho}\right)_y\right)_y \quad \text{in } G. \tag{4.5}$$

Similarly, we compute

$$\rho T_t = \rho T_{\underline{t}} + \rho T_y y_t = \rho T_{\underline{t}} + \rho^2 D_1\rho_y t_y$$

and

$$\begin{aligned} \rho T_t &= (\rho T)_t - \rho_t T = (\rho D_2 T_x)_x + D_1\rho_x T_x \\ &= \rho(\rho^2 D_2 T_y)_y + \rho^2 D_1\rho_y T_y. \end{aligned}$$

Thus, since $\rho > 0$, T satisfies

$$T_{\underline{t}} = (\rho^2 D_2 T_y)_y \quad \text{in } G. \tag{4.6}$$

We define

$$u(y, \underline{t}) \equiv \rho^{-1}(x, t), \qquad v(y, \underline{t}) \equiv T(x, t), \tag{4.7}$$

and

$$A(u, v) \equiv \rho^2 D_1(\rho, T), \qquad B(u, v) \equiv \rho^2 D_2(\rho, T). \tag{4.8}$$

By (4.5) and (4.6) and writing $t = \underline{t}$, we arrive at the problem

$$\text{(II)} \begin{cases} u_t = (A(u, v)u_y)_y & \text{in } G \\ v_t = (B(u, v)v_y)_y & \text{in } G \\ u_y(y, t) = v_y(y, t) = 0 & \text{for } y = 0, M; \quad 0 < t \leqq \tau \\ u(y, 0) = u_0(y);\ v(y, 0) = v_0(y) & \text{for } 0 < y < M, \end{cases}$$

where $u_0(y) \equiv \rho_0^{-1}(x)$ and $v_0(y) \equiv T_0(x)$.

Lemma 3.2 is a consequence of the following result.

LEMMA 4.1. *Let* $u_0,\ v_0 \in C^3([0, M])$ *and let* $s_\pm \in \mathbb{R}$ *be such that*

$$s_- \leqq u_0 \leqq S_+, \qquad s_- \leqq v_0 \leqq s_+ \quad \text{in } (0, M).$$

If $A,\ B \in C^2([s_-, s_+] \times [s_-, s_+])$ *and* $A,\ B > 0$ *on* $[s_-, s_+]^2$, *then problem* II *possesses a classical solution* $(u, v) \in \{C^{2,1}(\bar{G})\}^2$, *and*

$$s_- \leqq u \leqq s_+, \qquad s_- \leqq v \leqq s_+ \quad \text{in } G.$$

Remark. Lemma 4.1 is a consequence of general results about diagonal systems in divergence form [Am3]. Below we will sketch a more elementary proof.

Proof of Lemma 4.1. Let Y be the convex, closed subset of the Banach space $X = \{C(\bar{G})\}^2$, defined by

$$Y = \{(u, v) \in X;\ s_- \leqq u, v \leqq s_+\}.$$

We define a map $T: Y \to Y$ by

$$T(\tilde{u}, \tilde{v}) = (u, v),$$

where $u,\ v \in C(\bar{G}) \cap L^2(0, \tau; H^1((0, M)))$ are the (weak) solutions of the linear problems

$$\text{(III}_A\text{)} \begin{cases} u_t = (A(\tilde{u}, \tilde{v})u_y)_y & \text{in } G \\ u_y(0, t) = u_y(M, t) = 0 & \text{for } t \in (0, \tau] \\ u(y, 0) = u_0(y) & \text{for } y \in (0, M), \end{cases}$$

respectively,

$$\text{(III}_B\text{)} \begin{cases} v_t = (B(\tilde{u}, \tilde{v})v_y)_y & \text{in } G \\ v_y(0, t) = v_y(M, t) = 0 & \text{for } t \in (0, \tau] \\ v(0, y) = v_0(y) & \text{for } y \in (0, M) \end{cases}$$

(to prove that $T(\tilde{u}, \tilde{v}) \in Y$ we have to show that $s_- \leqq u,\ v \leqq s_+$ in G, but since u and v are only weak solutions this does not follow at once from the maximum principle; therefore we approximate u and v by classical solutions u_n and v_n which are obtained by replacing $A(\tilde{u}, \tilde{v})$ and $B(\tilde{u}, \tilde{v})$ by approximating smooth functions A_n and B_n; we leave the details to the reader).

By classical results [LSU] about linear equations,

$$\|u\|_{C^{\alpha,\alpha/2}(\bar{G})}, \|v\|_{C^{\alpha,\alpha/2}(\bar{G})} \leqq \mathscr{C} \quad \text{for } (\tilde{u}, \tilde{v}) \in Y \tag{4.9}$$

for some constants $\alpha \in (0, 1)$ and $\mathscr{C} > 0$ (to be more precise, (4.9) can be considered as an interior estimate if we extend u to the domain $[-M, 2M] \times [0, \tau]$ by reflecting u around the lines $y = 0$ and $y = M$; this interior estimate is given in [LSU, Chap. III, Thm. 10.1]). By (4.9) and the theorem of Arzelà-Ascoli, TY is compact, and, by Schauder's fixed point theorem [GT, Cor. 10.2], T has a fixed point in Y, which we denote by (u, v).

We can consider (u, v) as a weak solution of problem II. To prove that u and v have the required smoothness, it is enough to prove that $A(u, v)$ and $B(u, v)$ belong to $C^{1+\alpha,\alpha/2}(\bar{G})$, i.e., that

$$u, v \in C^{1+\alpha,\alpha/2}(\bar{G}). \tag{4.10}$$

We define $w: \bar{G} \to \mathbb{R}$ by

$$w(y, t) = \int_0^y u(s, t)\, ds. \tag{4.11}$$

Then w is a solution of the problem

$$w_t = A(u, v) w_{yy} \qquad \text{in } G$$

$$w(0, t) = 0 \quad \textit{and} \quad w(M, t) = \int_0^M u_0(y)\, dy \quad \text{for } t \in (0, \tau]$$

$$w(y, 0) = \int_0^y u_0(s)\, ds \qquad \text{for } y \in (0, M).$$

Since $A(u, v) \in C^{\alpha,\alpha/2}(\bar{G})$ for $\alpha \in (0, 1)$, it follows from the Schauder-type a priori estimates [LSU], [F] that $w \in C^{2+\alpha,1+\alpha/2}(\bar{G})$. Hence (4.10) follows for u. The proof for v is identical.

Remark 4.2. Except for the proof of (4.10), the proof of Lemma 4.1 can be generalized to arbitrary space dimensions. In particular, we obtain existence of weak continuous solutions of systems of the form

$$u_t = \operatorname{div}(A(u, v) \nabla u) \qquad \text{in } \Omega \times (0, \tau]$$

$$v_t = \operatorname{div}(B(u, v) \nabla v) \qquad \text{in } \Omega \times (0, \tau]$$

+boundary and initial conditions.

Remark 4.3. In this section we have used the transformation to Lagrangian coordinates to prove existence of solutions of problem I_ε. We always assumed that

$$\rho_\varepsilon(x, t) \geqq \varepsilon > 0 \quad \text{in } Q_{\varepsilon,\tau} \tag{4.12}$$

but it would be slightly misleading to suggest that we only need (4.12) to avoid the degeneracy of the parabolicity. Actually a condition like (4.12) is already necessary to make the transformation to Lagrangian coordinates work. Even in the case of the scalar heat equations (i.e., $D_1 \equiv 1$), we cannot relax it to, for example, the condition that $\rho > 0$ for $t > 0$. For details, we refer to [BdPU, § 4].

Finally we give the proof of Lemmas 3.4 and 3.5, announced in § 3.

Proof of Lemma 3.4. Let $(x_0, t_0) \in \mathscr{P} \cap Q_\tau$ and $B_r = \{x, t: (x-x_0)^2 + (t-t_0)^2 < r^2\}$. Let r be small enough so that $\rho \geqq 2\delta$ in B_r. Then

$$\rho_\varepsilon \geqq \delta \quad \text{in } B_r \tag{4.13}$$

for all ε small enough.

Step 1. The functions $y = y_\varepsilon$ are uniformly Hölder continuous in B_r. By (3.2) and (4.13) we have

$$\iint_{B_r} \left(\frac{\partial \rho_\varepsilon}{\partial x}\right)^2 dx\, dt \leqq \mathscr{C} \tag{4.14}$$

for some $\mathscr{C}$ which may depend on r but does not depend on ε. From (4.14) it follows that for $\{(x_1, x_2) \times (t_1, t_2)\} \subset B_r$, with $x_1 < x_2$ and $t_1 < t_2$

$$\begin{aligned} \left| \int_{x_1}^{x_2} y(x, t_1)\, dx - \int_{x_1}^{x_2} y(x, t_2)\, dx \right| &= \left| \int_{x_1}^{x_2} \int_{t_1}^{t_2} D_1 \rho_x\, dx\, dt \right| \\ &\leqq \mathscr{C} \left\{ \int_{x_1}^{x_2} \int_{t_1}^{t_2} (\rho_x)^2\, dx\, dt \right\}^{1/2} \cdot \sqrt{t_2 - t_1} \leqq \mathscr{C}_2 \sqrt{t_2 - t_1}. \end{aligned} \tag{4.15}$$

Moreover

$$\frac{\partial y_\varepsilon}{\partial x} = \rho_\varepsilon$$

and therefore

$$|y_\varepsilon(x + \Delta x, t) - y_\varepsilon(x, t)| \leqq \mathscr{C} |\Delta x|. \tag{4.16}$$

Using (4.15) and (4.16) we get that for every α small enough

$$\begin{aligned} |y_\varepsilon(x, t+\Delta t) - y_\varepsilon(x, t)| &\leqq \left| y_\varepsilon(x, t+\Delta t) - \frac{1}{\alpha} \int_x^{x+\alpha} y_\varepsilon(s, t+\Delta t)\, ds \right| \\ &\quad + \left| y_\varepsilon(x, t) - \frac{1}{\alpha} \int_x^{x+\alpha} y_\varepsilon(s, t)\, ds \right| \\ &\quad + \frac{1}{\alpha} \left| \int_x^{x+\alpha} [y_\varepsilon(s, t+\Delta t) - y_\varepsilon(s, t)]\, ds \right| \leqq \mathscr{C}\alpha + \mathscr{C} \frac{|\Delta t|^{1/2}}{\alpha}. \end{aligned}$$

Let $\alpha = |\Delta t|^{1/4}$. We obtain

$$|y_\varepsilon(x, t+\Delta t) - y_\varepsilon(x, t)| \leqq \mathscr{C} |\Delta t|^{1/4}. \tag{4.17}$$

The Hölder continuity follows from (4.16) and (4.17).

Step 2. Let $y_0 = y_\varepsilon(x_0, t_0)$, $\underline{t}_0 = t_0$ and $\mathrm{Im}\,(B_r)$ be the image of B_r in the y, $\underline{t}$ plane. Then $\mathrm{Im}\,(B_r)$ contains a rectangle $|y - y_0| < \beta$, $|\underline{t} - \underline{t}_0| < \Delta t$ where β and Δt do not depend on ε. For the proof we notice first that if $|x - x_0| < \tau/2$, $|t - t_0| < \tau/2$ then $(x, t) \in B_r$. Therefore for $|t - t_0| < \tau/2$ and for all x

$$|y(x, t) - y(x_0, t)| = \left| \int_{x_0}^{x} \rho(s, t)\, ds \right| \geqq \delta \cdot \min \left\{ |x - x_0|, \frac{r}{2} \right\}. \tag{4.18}$$

Let β be some positive number which we choose later and $\Delta t < \tau/2$ be so small that

$$|y(x_0, t) - y_0| < \beta \quad \text{for } |t - t_0| < \Delta t. \tag{4.19}$$

By Step 1 such Δt exists and does not depend on ε. Next we take the rectangle in y, $\underline{t}$ plane $K_{\beta,\varepsilon} = \{(y, t): |y - y_0| < \beta, |\underline{t} - t_0| < \Delta t\}$. We have using (4.18), (4.19) for $(y, t) \in K_{\beta,\varepsilon}$

$$\beta \geqq |y(x, t) - y_0| \geqq |y(x, t) - y(x_0, t)| - |y(x_0, t) - y_0|$$

$$\geqq \delta \cdot \min \{|x - x_0|, \tfrac{1}{2}r\} - \beta$$

and hence

$$2\beta \geqq \delta \min \left\{|x - x_0|, \frac{r}{2}\right\}.$$

Take $\beta < \frac{1}{4}\delta r$; then $\min \{|x - x_0|, \tau/2\} = |x - x_0|$ and therefore

$$|x - x_0| < \frac{2\beta}{\delta} < \frac{1}{2}r.$$

Finally $K_{\beta,\varepsilon} \subset \operatorname{Im}(B_\tau)$ and β, Δt do not depend on ε. Note that the last assertion means the equicontinuity of the inverse transformation $x = x(y, \underline{t})$, $t = \underline{t}$ at the point y_0, $\underline{t}_0$.

Step 3. We may now apply Theorem 10.1 [LSU, Chap. III] to the solution of uniformly parabolic equations (4.5), (4.6) in $K_{\beta,\varepsilon}$. As a result, we get that the functions $\rho_\varepsilon(y, \underline{t})$, $T_\varepsilon(y, \underline{t})$ are uniformly Hölder continuous in the interior domain $\tilde{K}_\varepsilon \Subset K_{\beta,\varepsilon}$, $\operatorname{dist}(\tilde{K}_\varepsilon, \partial K_{\beta,\varepsilon}) > 0$. Combining this result with Step 1, we obtain that ρ_ε, T_ε are locally uniformly Hölder continuous in the x, t variables near (x_0, t_0) and thus the functions T_ε are locally uniformly Hölder continuous in $\mathcal{P} \cap (0, \tau)$.

Moreover, proceeding as in the proof of (4.10), it follows that

(4.20) $\qquad \rho_\varepsilon$ and T_ε are uniformly bounded in $C^{2+\alpha,1+\alpha/2}_{\text{loc}}(\tilde{K}_\varepsilon)$

as functions of y, $\underline{t}$.

Step 4. To complete the proof we have to show that T_ε are locally uniformly continuous in $\mathcal{P}$ down to $t = 0$. This follows from the standard arguments, using the previous considerations and the equicontinuity of $T_{0,\varepsilon}$ in $\mathcal{P}_0$.

Proof of Lemma 3.5. We proceed as in the proof of Lemma 3.4 and introduce the sets $K_{\beta,\varepsilon}$ and $\tilde{K}_\varepsilon \Subset K_{\beta,\varepsilon}$. By Step 1 of the proof of Lemma 3.4 the ball $B_{\tilde{r}} = \{(x, t): (x - x_0)^2 + (t - t_0)^2 < \tilde{r}^2\}$ belongs to $\tilde{K}_\varepsilon$. In view of (4.20) it is enough to prove that

(4.21)
$$\frac{\partial y_\varepsilon}{\partial x} \text{ and } \frac{\partial y_\varepsilon}{\partial t} \text{ are uniformly bounded in, respectively,}$$

$$C^{1+\alpha,\alpha/2}(B_{\tilde{r}}) \text{ and } C^{\alpha,\alpha/2}(B_{\tilde{r}}).$$

But $Y_\varepsilon(x, t) = y_\varepsilon(x, t) - y_\varepsilon(x_0, t_0)$ are uniformly bounded in $B_{\tilde{r}}$ and satisfy the equation

$$\frac{\partial Y_\varepsilon}{\partial t} = D_1 \frac{\partial^2 Y_\varepsilon}{\partial x^2}$$

where $D_1 = D_1(\rho_\varepsilon(x, t), T_\varepsilon(x, t))$ are uniformly bounded in $C^{\alpha,\alpha/2}(B_{\tilde{r}})$. Hence, applying once more a priori estimates for linear uniformly parabolic equation [LSU], [F] we obtain (4.21). This completes the proof.

REFERENCES

[AL] H. W. ALT AND S. LUCKHAUS, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311-341.

[Am1] H. AMANN, *Quasilinear evolution equations and parabolic systems*, Trans. Amer. Math. Soc., 293 (1986), pp. 191-227.

[Am2] ———, *Dynamic theory of quasilinear parabolic equations.* II. *Reaction-diffusion systems*, to appear.

[Am3] ———, *Dynamic theory of quasilinear parabolic equations.* III. *Global existence*, in preparation.

[Ar] D. G. ARONSON, *The porous medium equation*, in Some Problems in Nonlinear Diffusion, A. Fasano and M. Primicerio, eds., Lecture Notes in Math. 1224, Springer-Verlag, Berlin, New York, 1986.

[BvDEZ] M. BERTSCH, C. J. VAN DUYN, J. R. ESTEBAN, AND ZHANG HONGFEI, *Regularity of the free boundary in a doubly degenerate parabolic equation*, Comm. Partial Differential Equations, 14 (1989), pp. 391-412.

[BdP] M. BERTSCH AND R. DAL PASSO, *A numerical treatment of a superdegenerate equation with applications to the porous medium equation*, Quart. Appl. Math., to appear.

[BdPU] M. BERTSCH, R. DAL PASSO, AND M. UGHI, *Nonuniqueness of solutions of a degenerate parabolic equation*, Ann. Mat. Pura Appl., to appear.

[BK1] M. BERTSCH AND S. KAMIN, *Some properties of degenerate parabolic equations*, in preparation.

[BK2] ———, *A system of degenerate parabolic equations from plasma physics: the large time behaviour*, in preparation.

[dB] E. DI BENEDETTO, *Continuity of weak solutions to a general porous medium equation*, Indiana Univ. Math. J., 32 (1983), pp. 83-118.

[F] A. FRIEDMAN, *Partial differential equations of parabolic type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[GT] D. GILBARG AND N. S. TRUDINGER, *Elliptic partial differential equations of second order*, Springer-Verlag, Berlin, New York, 1977.

[LSU] O. A. LADYZENSKAYA, V. A. SOLONNIKOV, AND N. N. URAL'TZEVA, *Linear and quasilinear equations of parabolic type*, Transl. Math. Monographs 23, American Mathematical Society, Providence, RI, 1968.

[RH1] P. ROSENAU AND J. M. HYMAN, *Analysis of nonlinear mass and energy diffusion*, Phys. Rev. A, 32 (1985), pp. 2370-2373.

[RH2] ———, *Plasma diffusion across a magnetic field*, Phys. D, 20 (1986), pp. 444-446.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 917-934, July 1990
© 1990 Society for Industrial and Applied Mathematics
006

AN INTEGRAL EQUATION METHOD FOR A PROBLEM WITH MIXED BOUNDARY CONDITIONS*

WILLIAM McLEAN†

Abstract. Consider the problem of finding a complex function $f = u + iv$, which is holomorphic in a given domain, with the real part u taking prescribed values on one part of the boundary, and the imaginary part v taking prescribed values on the remainder of the boundary. (This is essentially equivalent to solving Laplace's equation subject to mixed Dirichlet and Neumann boundary conditions.) By virtue of the Cauchy integral formula, the unknown boundary values of u and v satisfy a 2×2 system of singular integral equations, which can be solved in a certain class of weighted L_p spaces, by applying a result of I. Gohberg and N. Krupnik. The values of both u and v are then known over the whole of the boundary, and so f can be computed using the Cauchy integral formula.

Key words. mixed boundary conditions, singular integral equations

AMS(MOS) subject classifications. 45F15, 35C15

1. Introduction.

Let $0 < \alpha < 1$, and suppose Γ is a simple, closed $C^{1,\alpha}$ curve in the plane. (Thus, Γ satisfies the Lyapunov condition, see, e.g., Mikhlin [12, Chap. 18].) The complement, $\mathbb{R}^2 \backslash \Gamma$, consists of a bounded (interior) component Ω_i and an unbounded (exterior) component Ω_e. It is convenient to take the usual one-point compactification $\mathbb{R}^2 \cup \{\infty\}$ of the plane, and to allow $\infty \in \Omega_e$. In this way, both Ω_i and Ω_e are simply connected and have compact closure.

The curve Γ is given a counterclockwise orientation, and the forward unit tangent vector is denoted by τ. The unit normal vector ν is chosen to point into Ω_i so that, at every point of Γ, the ordered pair of vectors (τ, ν) is a right-hand basis for $\mathbb{R}^2$, i.e., the cross product $\tau \times \nu = (0, 0, 1) \in \mathbb{R}^3$. Also, to simplify notation, $\mathbb{R}^2$ will be freely identified with the complex plane $\mathbb{C}$ by writing a typical vector as $x = (x_1, x_2) = x_1 + ix_2$.

In what follows, we consider the problem of finding a complex function $f = u + iv$ which is holomorphic in Ω_i, and satisfies mixed boundary conditions of the form

$$u|_{\Gamma_1} = g|_{\Gamma_1}, \qquad v|_{\Gamma_2} = g|_{\Gamma_2}. \tag{1.1}$$

Here, Γ_1 and Γ_2 are nonoverlapping sub-arcs of Γ satisfying $\bar{\Gamma}_1 \cup \bar{\Gamma}_2 = \Gamma$, and g is a given function defined on Γ. A classical treatment of this problem for the case when Γ is the unit circle may be found in [14, § 94].

The requirement that f be holomorphic is equivalent to assuming u and v are conjugate harmonic functions, i.e., they satisfy the Cauchy-Riemann equations

$$\frac{\partial u}{\partial x_1} = \frac{\partial v}{\partial x_2}, \qquad \frac{\partial u}{\partial x_2} = -\frac{\partial v}{\partial x_1},$$

everywhere in Ω_i. Our problem could be formulated in terms of u alone. Indeed, provided u and v are sufficiently smooth at the boundary, the Cauchy-Riemann equations hold along Γ in the form

$$\frac{\partial u}{\partial \tau} = \frac{\partial v}{\partial \nu}, \qquad \frac{\partial u}{\partial \nu} = -\frac{\partial v}{\partial \tau}, \tag{1.2}$$

* Received by the editors December 23, 1987; accepted for publication (in revised form) August 8, 1989.
† School of Mathematics, University of New South Wales, Kensington 2033, Australia. This research was carried out while the author was a Queen Elizabeth II Fellow at the University of Tasmania.

and therefore, if we let $h = -dg/d\tau$, then u satisfies

$$\begin{aligned} \Delta u &= 0 \quad \text{on } \Omega_i, \\ u &= g \quad \text{on } \Gamma_1, \\ \frac{\partial u}{\partial \nu} &= h \quad \text{on } \Gamma_2, \end{aligned} \tag{1.3}$$

where $\Delta = (\partial/\partial x_1)^2 + (\partial/\partial x_2)^2$ is the Laplacian. Thus, we can think of u as a solution of the mixed Dirichlet and Neumann problem for the Laplace equation. Note that $h|_{\Gamma_2}$ determines $g|_{\Gamma_2}$ uniquely up to an additive constant, reflecting the fact that, given u, the harmonic conjugate v is unique up to a constant.

Before discussing our own approach to problem (1.3), we will briefly review some established results. For $j = 1$ and 2, let $H^{1/2}(\Gamma_j)$ denote the Sobolev space consisting of the restrictions to Γ_j of functions in $H^{1/2}(\Gamma)$, and let $\tilde{H}^{-1/2}(\Gamma_j)$ denote the dual of $H^{1/2}(\Gamma_j)$ (cf. Costabel and Stephan [3, pp. 178–179]).

Suppose $g|_{\Gamma_1} \in H^{1/2}(\Gamma_1)$ and $h|_{\Gamma_2} \in \tilde{H}^{-1/2}(\Gamma_2)$, and let

$$\mathcal{V} = \{\varphi \in H^1(\Omega_i) : \varphi|_{\Gamma_1} = 0\}.$$

By using the trace theorem to extend g to a function in $H^1(\Omega_i)$ and then putting $u = w + g$, we can reformulate (1.3) as the following variational problem: find $w \in \mathcal{V}$ satisfying

$$\int_{\Omega_i} \nabla w \cdot \nabla \varphi \, dx = -\int_{\Omega_i} \nabla g \cdot \nabla \varphi \, dx - \int_{\Gamma_2} h\varphi |dt| \quad \text{for all } \varphi \in \mathcal{V}. \tag{1.4}$$

(Throughout this paper, we use the notation $\nabla \varphi = (\partial\varphi/\partial x_1, \partial\varphi/\partial x_2)$ for the gradient of the function φ. Also, the arc-length measure on Γ is denoted by $|dt|$, since this is consistent with writing $\int_\Gamma f(t)\, dt$ for the contour integral of f along Γ.) The Lax–Milgram theorem implies that there exists a unique function $w \in \mathcal{V}$ satisfying (1.4), because the Dirichlet bilinear form is coercive on $\mathcal{V}$; see Ciarlet [1, p. 21]. Hence, there exists a unique variational solution $u \in H^1(\Omega_i)$ of the original mixed boundary value problem (1.3). The same analysis also works in the case when Γ is a polygon rather than a smooth curve; see [3, p. 181] and [7, Chap. 4].

In applications, we are usually interested in the variational solution, because this has finite energy, i.e., $\nabla u \in L_2(\Omega_i)$. Later, in Example 5.5, we will construct an infinite number of singular eigensolutions of the mixed problem, each of which is physically spurious in the sense that its gradient fails to be square-integrable near one or another of the so-called collision points, where Γ_1 and Γ_2 meet. This fact must be taken into account when discussing nonvariational solutions of (1.3), since it obviously affects the question of uniqueness.

If the function u is harmonic on Ω_i and, say, is in $C^2(\bar{\Omega}_i)$, then a well-known argument [8], [16] involving Green's second identity implies

$$u(z) = \frac{1}{2\pi} \int_\Gamma \frac{\nu(t) \cdot (z-t)}{|z-t|^2} u(t) |dt| - \frac{1}{2\pi} \int_\Gamma \log\left(\frac{1}{|z-t|}\right) \frac{\partial u}{\partial \nu}(t) |dt| \tag{1.5}$$

for $z \in \Omega_i$, and

$$u(z) = \frac{1}{\pi} \int_\Gamma \frac{\nu(t) \cdot (z-t)}{|z-t|^2} u(t) |dt| - \frac{1}{\pi} \int_\Gamma \log\left(\frac{1}{|z-t|}\right) \frac{\partial u}{\partial \nu}(t) |dt| \tag{1.6}$$

for $z \in \Gamma$. More generally, it can be shown [3, pp. 183, 208] that these formulae are valid for any variational solution u of (1.3). If the boundary conditions from (1.3) are

substituted into (1.6), then we obtain a 2×2 system of integral equations for the unknown functions $\partial u/\partial \nu|_{\Gamma_1}$ and $u|_{\Gamma_2}$. This system determines an invertible linear operator [3, p. 212]

$$\left(\left.\frac{\partial u}{\partial \nu}\right|_{\Gamma_1}, u|_{\Gamma_2}\right) \mapsto (g|_{\Gamma_1}, h|_{\Gamma_2})$$

$$\tilde{H}^{-1/2}(\Gamma_1)\times H^{1/2}(\Gamma_2) \to H^{1/2}(\Gamma_1)\times \tilde{H}^{-1/2}(\Gamma_2),$$

provided Γ has the following property: the homogeneous, first-kind integral equation

$$\int_\Gamma \log\left(\frac{1}{|z-t|}\right)\varphi(t)|dt| = 0, \qquad z\in\Gamma,$$

has only the trivial solution $\varphi(t)\equiv 0$. (This property fails to hold only when Γ is "exceptional" in a certain sense [8, p. 52], [10, p. 142].) Finally, if the solutions of the boundary integral equations, together with the boundary conditions from (1.3), are substituted into the right-hand side of (1.5), then the variational solution is recovered [3, p. 213].

In this paper, we will solve (1.3) using an integral equation method different from the standard one just described. The Cauchy integral formula will take the place of (1.5) and, by applying boundary conditions (1.1), a 2×2 system of singular integral equations will be obtained for the unknown functions $v|_{\Gamma_1}$ and $u|_{\Gamma_2}$. This system will be solved in weighted L_p spaces by applying a theorem of Gohberg and Krupnik [6]. Our approach avoids the use of half-order Sobolev spaces, but the price of this simplification is the severing of a direct connection with the variational solution.

Now for an overview of the paper. In § 2, a result on L_p boundary values of harmonic functions is stated, and a class of weighted L_p spaces is introduced, along with some miscellaneous notation and terminology. Following this, in § 3, we summarize a large number of results from the L_p theory of layer potentials and the Cauchy integral—material that is referred to many times in subsequent sections. Next comes a very brief section devoted solely to describing the result of Gohberg and Krupnik [6] concerning the index of a singular integral operator with piecewise-continuous, matrix-valued coefficients. The fifth—and longest—section treats the integral equation method itself, with the main results on existence and uniqueness of solutions in weighted L_p spaces being stated as Theorem 5.4. The final section is devoted to the analogous mixed boundary value problem for f holomorphic in the unbounded region Ω_e. Here some additional complications arise, connected with the value of f at infinity.

2. Preliminaries. The scalar product in $\mathbb{R}^2$ will be written

$$x\cdot y = x_1y_1 + x_2y_2 = \Re(\bar{x}y),$$

where $x=(x_1, x_2)$ and $y=(y_1, y_2)$.

In order to discuss nontangential limits on Γ of functions defined on Ω_i and/or Ω_e, fix an angle θ such that $0<\theta<\pi/2$, and a number $\varepsilon>0$. For $t\in\Gamma$, define the *interior cone* with vertex t, axis $\nu(t)$, half-angle θ and height ε, by

$$C_i(t) = \{x\in\mathbb{R}^2: |x-t|\cos\theta < \nu(t)\cdot(x-t) < \varepsilon\}.$$

Similarly, define the *exterior cone*

$$C_e(t) = \{x\in\mathbb{R}^2: |x-t|\cos\theta < -\nu(t)\cdot(x-t) < \varepsilon\}.$$

It is assumed ε is chosen small enough to ensure that the corresponding interior and exterior cones with height 2ε are contained in Ω_i and Ω_e, respectively.

DEFINITION 2.1. If the function u is defined on Ω_i, then the interior nontangential maximal function $u_i^*: \Gamma \to [0, \infty]$ is defined by

$$u_i^*(t) = \sup_{x \in C_i(t)} |u(x)|, \qquad t \in \Gamma,$$

and the interior nontangential limit of u at t is denoted by

$$u_i(t) = \lim_{\substack{x \to t \\ x \in C_i(t)}} u(x), \qquad t \in \Gamma,$$

whenever this limit exists.

If u is defined on Ω_e instead of on Ω_i, then we define u_e^* and u_e in the obvious way, using exterior cones instead of interior cones. We also define the two-sided nontangential maximal function $u^* = \max\{u_i^*, u_e^*\}$.

The significance of these notions can be seen from the following result.

THEOREM 2.2 [4], [9]. *Suppose* $1 < p \leqq \infty$. *If* u *is harmonic on* Ω_i *and if* $u_i^* \in L_p(\Gamma)$, *then* $u_i(t)$ *exists for almost all* $t \in \Gamma$. *Moreover,* $u_i \in L_p(\Gamma)$, *and* u *equals the Poisson integral of* u_i.

We will say that u is harmonic on Ω_e, if u is harmonic on $\Omega_e \backslash \{\infty\}$ in the usual sense, and if, in addition, u is bounded at ∞. With this convention, Theorem 2.2 remains true for the exterior case. Indeed, if u is harmonic on Ω_e, then the function $z \mapsto u(1/z)$ is bounded and harmonic on a punctured neighbourhood of the origin, and therefore has a removable singularity at $z = 0$. This observation also shows that u admits a single-valued harmonic conjugate on Ω_e.

From this point onwards, we will always assume implicitly that the number p satisfies $1 < p < \infty$. The norm in $L_p(\Gamma)$ will be written

$$\|f\|_p = \left(\int_\Gamma |f(t)|^p |dt| \right)^{1/p},$$

and, more generally, we will introduce a class of weighted L_p norms as follows. Fix distinct points $t_{(1)}, \cdots, t_{(r)}$ lying on Γ, and fix real numbers $\beta_1, \cdots, \beta_r$ satisfying

$$-\frac{1}{p} < \beta_j < \frac{1}{q} \quad \text{for } 1 \leqq j \leqq r, \tag{2.1}$$

where q is the conjugate exponent to p, given by $1/p + 1/q = 1$. Define the weight function

$$\rho(t) = \prod_{j=1}^{r} |t - t_{(j)}|^{\beta_j}, \qquad t \in \Gamma, \tag{2.2}$$

and define the space $L_p(\Gamma, \rho)$ by putting

$$\|f\|_{L_p(\Gamma,\rho)} = \|\rho f\|_p = \left(\int_\Gamma |\rho(t) f(t)|^p |dt| \right)^{1/p}.$$

As is easily verified, condition (2.1) implies that there exist p_0 and p_1 such that $1 < p_0 \leqq p \leqq p_1 < \infty$ and

$$L_{p_1}(\Gamma) \subset L_p(\Gamma, \rho) \subset L_{p_0}(\Gamma),$$

with the inclusions being continuous. Furthermore, it is not difficult to show [13, p. 53] that $L_q(\Gamma, \rho^{-1})$ is the topological dual of $L_p(\Gamma, \rho)$ with respect to the usual pairing

$$\langle g, f \rangle = \int_\Gamma g(t) f(t) |dt|. \tag{2.3}$$

Obviously, $L_p(\Gamma, \rho) = L_p(\Gamma)$ when $\beta_1 = \cdots = \beta_r = 0$, i.e., when $\rho = 1$.

Given a bounded linear operator $A: L_p(\Gamma, \rho) \to L_p(\Gamma, \rho)$, its transpose is the bounded linear operator $A^t: L_q(\Gamma, \rho^{-1}) \to L_q(\Gamma, \rho^{-1})$ defined by

$$\langle A^t g, f\rangle = \langle g, Af\rangle, \qquad g \in L_q(\Gamma, \rho^{-1}), \quad f \in L_p(\Gamma, \rho).$$

The *nullity* $n(A)$ and *deficiency* $d(A)$ of A are defined by

$$\begin{aligned} n(A) &= \dim \ker A, \\ d(A) &= \dim L_p(\Gamma, \rho)/\overline{\operatorname{im} A}, \end{aligned} \tag{2.4}$$

where

$$\ker A = \{u \in L_p(\Gamma, \rho): Au = 0\},$$

$$\operatorname{im} A = \{f \in L_p(\Gamma, \rho): f = Au \text{ for some } u \in L_p(\Gamma, \rho)\}$$

are the kernel and image of A, respectively. If $\operatorname{im} A$ is closed, and if both $n(A)$ and $d(A)$ are finite, then A is said to be a *Fredholm* operator (or, sometimes, a *Noether* operator), and the integer

$$\operatorname{ind} A = n(A) - d(A) \tag{2.5}$$

is called the (analytical) *index* of A. In this case, A^t is a Fredholm operator on $L_q(\Gamma, \rho^{-1})$, and [11, p. 42]

$$n(A) = d(A^t), \qquad d(A) = n(A^t), \tag{2.6}$$

so $\operatorname{ind} A^t = -\operatorname{ind} A$.

3. Operators and potentials associated with the Cauchy integral. In this section, we fix some notation, as well as gathering together certain established results which will be used repeatedly in later sections. Recall that it is assumed $1 < p < \infty$.

The starting point for our considerations is the Cauchy integral formula

$$f(z) = \frac{1}{2\pi i} \int_\Gamma \frac{f(t)}{t - z}\, dt, \qquad z \in \Omega_i,$$

which is valid if, e.g., the function f is holomorphic in a neighbourhood of $\bar{\Omega}_i$. An elementary calculation shows

$$\frac{1}{2\pi i}\frac{dt}{t-z} = \frac{1}{2\pi}\left\{\frac{\nu(t)\cdot(z-t)}{|z-t|^2} + i\frac{\tau(t)\cdot(z-t)}{|z-t|^2}\right\}|dt|, \tag{3.1}$$

and this equation provides part of the motivation for the following definition.

DEFINITION 3.1. Suppose $\varphi \in L_p(\Gamma)$. For $z \notin \Gamma$, let

$$U\varphi(z) = \frac{1}{2\pi}\int_\Gamma \frac{\tau(t)\cdot(z-t)}{|z-t|^2}\varphi(t)|dt|,$$

$$V\varphi(z) = \frac{1}{2\pi}\int_\Gamma \log\left(\frac{1}{|z-t|}\right)\varphi(t)|dt|,$$

$$W\varphi(z) = \frac{1}{2\pi}\int_\Gamma \frac{\nu(t)\cdot(z-t)}{|z-t|^2}\varphi(t)|dt|,$$

and for $z \in \Gamma$, let

$$R\varphi(z) = \frac{1}{\pi} \int\!\!\!\!\!\!- _\Gamma \frac{\tau(t) \cdot (z-t)}{|z-t|^2} \varphi(t)|dt|,$$

$$S\varphi(z) = \frac{1}{i\pi} \int\!\!\!\!\!\!- _\Gamma \frac{\varphi(t)}{t-z}\, dt,$$

$$T\varphi(z) = \frac{1}{\pi} \int_\Gamma \frac{\nu(t) \cdot (z-t)}{|z-t|^2} \varphi(t)|dt|.$$

In the definitions of R and S, the slash through the integral sign indicates that the Cauchy principal value must be taken. By contrast, the operator T is only weakly singular, because

$$\nu(t) \cdot (z-t) = O(|z-t|^{1+\alpha}) \quad \text{for } z, t \in \Gamma.$$

This bound is a simple consequence of the assumption that Γ is $C^{1,\alpha}$.

Both $U\varphi$ and $W\varphi$ are harmonic on $\Omega_i \cup \Omega_e$. Also, $V\varphi$ is harmonic on $\mathbb{R}^2 \backslash \Gamma$, but since

$$V\varphi(z) = \left(\frac{1}{2\pi} \int_\Gamma \varphi(t)|dt| \right) \log \frac{1}{|z|} + O(z^{-1}) \quad \text{as } z \to \infty, \tag{3.2}$$

it follows that $V\varphi$ is bounded at ∞ if and only if $\int_\Gamma \varphi(t)|dt| = 0$. Note that the gradient of $V\varphi$ is

$$\nabla V\varphi(z) = \frac{1}{2\pi} \int_\Gamma \frac{t-z}{|t-z|^2} \varphi(t)|dt|, \qquad z \notin \Gamma.$$

Traditionally, $V\varphi$ and $W\varphi$ are called, respectively, the *single* and *double layer potentials* of φ.

The transposes of the operators R and T with respect to the pairing (2.3) are, formally,

$$R^t\varphi(z) = \frac{1}{\pi} \int\!\!\!\!\!\!- _\Gamma \frac{\tau(z) \cdot (t-z)}{|-z|^2} \varphi(t)|dt|,$$

$$T^t\varphi(z) = \frac{1}{\pi} \int_\Gamma \frac{\nu(z) \cdot (t-z)}{|t-z|^2} \varphi(t)|dt|,$$

where $z \in \Gamma$, and it follows from (3.1) that R, S, and T are related by

$$S = T + iR. \tag{3.3}$$

The next theorem gives the interior and exterior boundary values in $L_p(\Gamma)$ of the potentials introduced above. As usual, $C(\Gamma)$ denotes the Banach space of continuous functions on Γ, equipped with the maximum norm.

THEOREM 3.2 [5], [9], [13]. *Suppose $\varphi \in L_p(\Gamma)$. The bounds*

$$\|(U\varphi)^*\|_p \leqq c\|\varphi\|_p,$$

$$\|(\nabla V\varphi)^*\|_p \leqq c\|\varphi\|_p,$$

$$\|(W\varphi)^*\|_p \leqq c\|\varphi\|_p$$

hold, and $V\varphi$ is continuous across Γ, with

$$\|V\varphi\|_{C(\Gamma)} \leqq c\|\varphi\|_p. \tag{3.4}$$

The operators

$$R, S: L_p(\Gamma, \rho) \to L_p(\Gamma, \rho)$$

are bounded, the operator

$$T: L_p(\Gamma, \rho) \to L_p(\Gamma, \rho)$$

is compact, and the relations

$$\begin{aligned}
&(U\varphi)_i = \tfrac{1}{2}R\varphi = (U\varphi)_e,\\
&\tau \cdot (\nabla V\varphi)_i = \tfrac{1}{2}R^t\varphi = \tau \cdot (\nabla V\varphi)_e,\\
&(W\varphi)_i = \tfrac{1}{2}(I+T)\varphi,\\
&\nu \cdot (\nabla V\varphi)_e = \tfrac{1}{2}(I+T^t)\varphi,\\
&(W\varphi)_e = \tfrac{1}{2}(-I+T)\varphi,\\
&\nu \cdot (\nabla V\varphi)_i = \tfrac{1}{2}(-I+T^t)\varphi
\end{aligned}$$

are valid, where I denotes the identity operator. Finally, the Cauchy integral

$$\Phi(z) = \frac{1}{2\pi i}\int_\Gamma \frac{\varphi(t)}{t-z}\,dt, \qquad z \notin \Gamma, \tag{3.5}$$

satisfies the estimate $\|\Phi^*\|_p \leqq c\|\varphi\|_p$, *as well as the Sokhotski–Plemelj formulae*

$$\begin{aligned}
&\Phi_i = \tfrac{1}{2}(I+S)\varphi,\\
&\Phi_e = \tfrac{1}{2}(-I+S)\varphi.
\end{aligned} \tag{3.6}$$

Here, the generic constant $c > 0$ is independent of φ, but is dependent, e.g., on the vertex angle θ of the interior and exterior cones used to define the nontangential maximal functions.

The following technical lemma will help in establishing further properties of the operators appearing in the preceding theorem and will be used again in §§ 5 and 6.

LEMMA 3.3. *Let* $\varphi \in L_p(\Gamma)$. *The single layer potential* $u = V\varphi$ *satisfies*

$$\int_{\Omega_i} |\nabla u|^2\,dx = -\int_\Gamma u_i\left(\frac{\partial u}{\partial \nu}\right)_i |dt|. \tag{3.7}$$

Moreover, if $\int_\Gamma \varphi |dt| = 0$, *then*

$$\int_{\Omega_e} |\nabla u|^2\,dx = \int_\Gamma u_e\left(\frac{\partial u}{\partial \nu}\right)_e |dt|. \tag{3.8}$$

Proof. First suppose that φ is Hölder continuous on Γ, so that ∇u has unique continuous extensions from Ω_i to $\bar{\Omega}_i$, and from Ω_e to $\bar{\Omega}_e$ (see, e.g., Smirnov [15, p. 594]). Since u is harmonic,

$$\nabla \cdot (u\nabla u) = |\nabla u|^2 \quad \text{on } \mathbb{R}^2 \backslash \Gamma,$$

and thus, by the divergence theorem,

$$\int_{\Omega_i} |\nabla u|^2\,dx = \int_\Gamma (-\nu) \cdot (u\nabla u)_i |dt| = -\int_\Gamma u_i\left(\frac{\partial u}{\partial \nu}\right)_i |dt|,$$

which establishes (3.7).

Introduce $\Omega_r = \{x \in \Omega_e\colon |x| < r\}$ and $\Gamma_r = \{x \in \mathbb{R}^2\colon |x| = r\}$ for all r sufficiently large; then, once again, the divergence theorem implies

$$\int_{\Omega_r} |\nabla u|^2\,dx = \int_\Gamma u_e\left(\frac{\partial u}{\partial \nu}\right)_e |dt| + \int_{\Gamma_r} u\frac{\partial u}{\partial \nu}|dt|, \tag{3.9}$$

with ν on Γ_r being the outward unit normal to Ω_r. If $\int_\Gamma \varphi|dt| = 0$, then, by (3.2), we have $u(x) = O(|x|^{-1})$ and $\nabla u(x) = O(|x|^{-2})$ as $|x| \to \infty$; hence $\int_{\Gamma_r} u\, \partial u/\partial \nu\, |dt| = O(r^{-2})$ and (3.8) follows from (3.9) by sending $r \to \infty$. (Note that if $\int_\Gamma \varphi(t)|dt| \neq 0$, then $u(x) = O(\log|x|)$ and $\nabla u(x) = O(|x|^{-1})$ as $|x| \to \infty$, so $\int_{\Gamma_r} u\, \partial u/\partial \nu\, |dt| = O(\log r)$ as $r \to \infty$.)

Now consider the general case, where $\varphi \in L_p(\Gamma)$. Since $u = V\varphi$ is continuous on Γ, and since Γ is $C^{1,\alpha}$, it can be shown [10, p. 136] that

$$\int_{\Omega_i \cup \Omega_e} |\nabla u|^2\, dx = \int_\Gamma \varphi u |dt|.$$

By Hölder's inequality and (3.4),

$$\left| \int_\Gamma \varphi u |dt| \right| \leqq \|\varphi\|_p \|u\|_q \leqq c(\|\varphi\|_p)^2,$$

and therefore

$$\left(\int_{\Omega_i \cup \Omega_e} |\nabla u|^2\, dx \right)^{1/2} \leqq c\|\varphi\|_p. \tag{3.10}$$

Choose a sequence of Hölder continuous functions φ_n converging to φ in $L_p(\Gamma)$. Put $u_n = V\varphi_n$; then, as we have just seen,

$$\int_{\Omega_i} |\nabla u_n|^2\, dx = -\int_\Gamma (u_n)_i \left(\frac{\partial u_n}{\partial \nu} \right)_i |dt|. \tag{3.11}$$

Replacing φ by $\varphi_n - \varphi$ in (3.4) and (3.10), it follows that $u_n \to u$ in $C(\Gamma)$ and $\nabla u_n \to \nabla u$ in $L_2(\Omega_i)$. Moreover, Theorem 3.2 implies

$$(\partial u_n/\partial \nu)_i = \tfrac{1}{2}(-I + T')\varphi_n \to (\partial u/\partial \nu)_i = \tfrac{1}{2}(-I + T')\varphi$$

in $L_p(\Gamma)$, and therefore (3.7) follows by sending $n \to \infty$ in (3.11). A similar argument can be used to prove (3.8), since if $\int_\Gamma \varphi(t)|dt| = 0$, then the Hölder continuous functions φ_n can be chosen in such a way that $\int_\Gamma \varphi_n(t)|dt| = 0$ for every n. □

The results of the next theorem are all well established, but there does not appear to be a single convenient reference for them. Thus, most of the proof is outlined.

THEOREM 3.4. *The identities*

$$S^2 = I, \quad R^2 = T^2 - I, \quad RT + TR = 0, \tag{3.12}$$

hold on $L_p(\Gamma, \rho)$, *and the operators*

$$S, I + T : L_p(\Gamma, \rho) \to L_p(\Gamma, \rho)$$

are invertible. Moreover,

$$R, I - T : L_p(\Gamma, \rho) \to L_p(\Gamma, \rho)$$

are Fredholm operators with zero index, and

$$\ker R = \ker (I - T) = \operatorname{span}\{1\}.$$

Finally, there is a real-valued function $\psi \in L_p(\Gamma)$ *such that*

$$\ker R^t = \ker (I - T^t) = \operatorname{span}\{\psi\}$$

and

$$\int_\Gamma \psi(t)|dt| = 1. \tag{3.13}$$

Proof. The books [6] and [13, pp. 46, 51] both contain proofs that $S^2 = I$ on $L_p(\Gamma, \rho)$, and (3.3) implies

$$S^2 = (T^2 - R^2) + i(TR + RT),$$

so the second two identities in (3.12) follow immediately from the first one. Obviously, S is invertible with $S^{-1} = S$, and the operators $I \pm T$ are Fredholm with zero index because T is compact. Similarly, R is Fredholm with zero index because, by (3.3), it differs from the invertible operator $-iS$ by the compact operator iT.

To prove the remaining assertions, we begin by observing that if $\varphi = 1$, then the Cauchy integral (3.5) satisfies $\Phi(z) = 1$ for all $z \in \Omega_i$, and hence the first of the Sokhotski-Plemelj formulae (3.6) implies $S1 = 1$. Thus, it follows from (3.3) that $1 \in \ker(I - T)$ and $1 \in \ker R$.

Next, we claim $\ker(I + T^t) = \{0\}$, which implies $I + T$ is invertible since $\operatorname{ind}(I + T) = 0$. Let $\varphi \in \ker(I + T^t)$; then $\int_\Gamma \varphi |dt| = \langle \varphi, 1\rangle = \langle -T^t\varphi, 1\rangle = -\langle \varphi, T1\rangle = -\langle \varphi, 1\rangle = -\int_\Gamma \varphi |dt|$, and hence $\int_\Gamma \varphi |dt| = 0$. Thus, if $u = V\varphi$, then the second part of Lemma 3.3 shows $\nabla u = 0$ on Ω_e, since $(\partial u/\partial \nu)_e = \frac{1}{2}(I + T^t)\varphi = 0$. Therefore, u is constant on Ω_e, and, since $u(\infty) = 0$ by (3.2), we conclude $u = 0$ on Ω_e. Hence, $u_i = u_e = 0$ on Γ, and so, by the maximum principle, $u = 0$ on Ω_i, implying $\varphi = (\partial u/\partial \nu)_e - (\partial u/\partial \nu)_i = 0$.

Suppose $\varphi \in \ker(I - T^t)$ and $\int_\Gamma \varphi |dt| = 0$. If $u = V\varphi$, then Lemma 3.3 implies $\nabla u = 0$ on Ω_i, since $(\partial u/\partial \nu)_i = -\frac{1}{2}(I - T^t)\varphi = 0$. Thus, $u_e = u_i$ is constant on Γ, and since $u(z) = O(z^{-1})$ as $|z| \to \infty$, it follows by the maximum principle that $u = 0$ everywhere in $\mathbb{R}^2$. Hence, $\varphi = (\partial u/\partial \nu)_e - (\partial u/\partial \nu)_i = 0$, and using this observation, we will now show $n(I - T^t) \leqq 1$—recall the notation in (2.4). Let φ_1 and φ_2 be nontrivial elements of $\ker(I - T^t)$; then $\int_\Gamma \varphi_1 |dt| \neq 0$ and $\int_\Gamma \varphi_2 |dt| \neq 0$, so there exist nonzero scalars λ_1 and λ_2 such that the function $\varphi = \lambda_1\varphi_1 + \lambda_2\varphi_2 \in \ker(I - T^t)$ satisfies $\int_\Gamma \varphi |dt| = 0$. Arguing as before, we see $\varphi = 0$, which means φ_1 and φ_2 are linearly dependent.

Remembering $1 \in \ker(I - T)$ and $\operatorname{ind}(I - T) = 0$, it follows that $1 \leqq n(I - T) = n(I - T^t) \leqq 1$, and consequently $\ker(I - T) = \operatorname{span}\{1\}$ and $\ker(I - T^t) = \operatorname{span}\{\psi\}$, for some function $\psi \in L_p(\Gamma)$ which is not identically zero. Moreover, $\int_\Gamma \psi |dt| \neq 0$, as otherwise the argument of the previous paragraph would show $\psi = 0$; hence we may assume (3.13) holds.

Finally, suppose $\varphi \in \ker R$; then $0 = R^2\varphi = (I + T)(I - T)\varphi$ and so $(I - T)\varphi = 0$, which shows $\ker R \subset \ker(I - T)$. Also, $\psi \in \ker R^t$, because if $u = V\psi$, then, arguing as before, u is constant on Ω_i and hence $R^t\psi = 2(\partial u/\partial \tau)_i = 0$. Since $\operatorname{ind} R = 0$, we see $1 \leqq n(R^t) = n(R) \leqq n(I - T) = 1$, and thus $\ker R^t = \operatorname{span}\{\psi\}$ and $\ker R = \operatorname{span}\{1\}$. □

4. Systems of singular integral equations with piecewise continuous coefficients. We will describe a result of Gohberg and Krupnik [6], which is fundamental to the analysis of the integral equation method in §§ 5 and 6.

Let $PC^{n\times n}(\Gamma)$ denote the set of piecewise continuous functions defined on Γ and taking values in $\mathbb{C}^{n\times n}$, the set of $n \times n$ complex matrices. Thus, $a \in PC^{n\times n}(\Gamma)$ if $a : \Gamma \to \mathbb{C}^{n\times n}$ is continuous except at finitely many points, and if, for each point of discontinuity $x \in \Gamma$, both one-sided limits

$$a(x \pm 0) = \lim_{\substack{t \to x^{\pm} \\ t \in \Gamma}} a(t)$$

exist. Here, the + direction is that of the forward unit tangent vector τ. With each function $a \in PC^{n\times n}(\Gamma)$, we associate a continuous, closed, oriented curve $C_{p,\rho}(a)$ as follows.

First, for $0<\delta<2\pi$, define the function $c_\delta:[0,1]\to\mathbf{C}$ by

$$c_\delta(\mu)=\begin{cases}\dfrac{e^{i(\pi-\delta)\mu}\sin(\pi-\delta)\mu}{e^{i(\pi-\delta)}\sin(\pi-\delta)}, & \delta\neq\pi,\\ \mu, & \delta=\pi.\end{cases}$$

This function parametrizes a circular arc or, in the case $\delta=\pi$, a straight line, beginning at zero and ending at 1. When $0<\delta<\pi$, the arc coincides with the locus of points in the lower half-plane, at which the interval $[0,1]$ subtends an angle δ. Otherwise, i.e., when $\pi<\delta<2\pi$, the arc consists of the points in the upper half-plane at which the angle subtended by $[0,1]$ is $2\pi-\delta$.

Let $t_{(1)},\cdots,t_{(r)}$ and $\beta_1,\cdots,\beta_r$ be as in the definition (2.2) of the weight function ρ, and put

$$\delta(t)=\begin{cases}2\pi/p & \text{if } t\in\Gamma\backslash\{t_{(1)},\cdots,t_{(r)}\},\\ 2\pi(\beta_j+1/p) & \text{if } t=t_{(j)} \text{ for some } j\in\{1,\cdots,r\}.\end{cases}$$

Given $a\in PC^{n\times n}(\Gamma)$, define $a_{p,\rho}:\Gamma\times[0,1]\to\mathbb{C}^{n\times n}$ by

$$a_{p,\rho}(t,\mu)=(1-c_{\delta(t)}(\mu))a(t-0)+c_{\delta(t)}(\mu)a(t+0),\qquad t\in\Gamma,\quad 0\leqq\mu\leqq 1,$$

and observe that $a_{p,\rho}(t,\mu)=a(t)$ whenever a is continuous at t. The curve $C_{p,\rho}(a)$, which was mentioned earlier, is defined as the range of the function $(t,\mu)\mapsto\det a_{p,\rho}(t,\mu)$, where $(t,\mu)\in\Gamma\times[0,1]$. Note that the orientation of Γ induces an orientation of $C_{p,\rho}(a)$ in a natural way. If $C_{p,\rho}(a)$ does not pass through the origin, then the function a is said to be *p,ρ-regular*, in which case we can define

$$\operatorname{ind}_{p,\rho}(a)=\text{winding number of } C_{p,\rho}(a) \text{ about } 0.$$

This integer is called the *p,ρ-index* of a.

Next, we introduce the operators

$$P=\tfrac{1}{2}(I+S),\qquad Q=\tfrac{1}{2}(I-S),$$

which obviously satisfy

$$P+Q=I,\qquad P-Q=S. \tag{4.1}$$

Furthermore, because $S^2=I$, it follows that P and Q are projections, i.e., $P^2=P$ and $Q^2=Q$. Finally, we define the $n\times n$ diagonal operators $\mathbf{P}=\operatorname{diag}(P,\cdots,P)$ and $\mathbf{Q}=\operatorname{diag}(Q,\cdots,Q)$, which are bounded on the n-fold product space $L_p^n(\Gamma,\rho)=L_p(\Gamma,\rho)\times\cdots\times L_p(\Gamma,\rho)$.

THEOREM 4.1 [6], [13, p. 129]. *Suppose $a,b\in PC^{n\times n}(\Gamma)$. The operator*

$$A=a\mathbf{P}+b\mathbf{Q}:L_p^n(\Gamma,\rho)\to L_p^n(\Gamma,\rho)$$

is Fredholm if and only if $b^{-1}a$ exists and is a p,ρ-regular function in $PC^{n\times n}(\Gamma)$. In this case, $\operatorname{ind}A=-\operatorname{ind}_{p,\rho}(b^{-1}a)$.

Actually, the conclusions of Theorem 4.1 are valid even when Γ is only piecewise $C^{1,\alpha}$, i.e., when Γ is permitted to have finitely many corners (but no cusps).

5. The interior problem. Let $t_{(1)}$ and $t_{(2)}=t_{(0)}$ be two distinct points lying on Γ. We think of $t_{(0)}$ as the starting point and $t_{(2)}$ as the finishing point of Γ, as it is traversed in the counterclockwise sense. For $j=1$ and 2, let Γ_j denote the open sub-arc of Γ from $t_{(j-1)}$ to $t_{(j)}$, so that

$$\Gamma=\Gamma_1\cup\{t_{(1)}\}\cup\Gamma_2\cup\{t_{(2)}\}$$

is a disjoint union, and, as in the Introduction, $\bar{\Gamma}_1 \cup \bar{\Gamma}_2 = \Gamma$. The notation of § 4 applies with $r=2$, and thus the weight function ρ is given by

$$\rho(t) = |t - t_{(1)}|^{\beta_1} |t - t_{(2)}|^{\beta_2}, \qquad t \in \Gamma. \tag{5.1}$$

Guided by Theorem 2.2, a precise statement of the interior problem, described informally in § 1, is as follows:

P_i Given $g \in L_p(\Gamma)$, find a function $f = u + iv$ which is holomorphic in Ω_i, has its (interior) nontangential maximal function $f_i^* \in L_p(\Gamma)$, and satisfies the mixed boundary conditions

$$u_i|_{\Gamma_1} = g|_{\Gamma_1}, \qquad v_i|_{\Gamma_2} = g|_{\Gamma_2}.$$

The results of this section will imply that P_i has a unique solution when $2 < p < \infty$. First, we investigate more closely the Cauchy integral formula.

THEOREM 5.1. *If $f = u + iv$ is holomorphic in Ω_i, and if $f_i^* \in L_p(\Gamma)$, then*

$$f(z) = \frac{1}{2\pi i} \int_\Gamma \frac{f_i(t)}{t - z}\, dt, \qquad z \in \Omega_i, \tag{5.2}$$

and $Pf_i = f_i$, or, equivalently,

$$\left.\begin{aligned} u &= Wu_i - Uv_i \\ v &= Wv_i + Uu_i \end{aligned}\right\} \quad \text{on } \Omega_i, \tag{5.3}$$

and

$$(I - T)u_i = -Rv_i, \tag{5.4}$$

$$(I - T)v_i = Ru_i. \tag{5.5}$$

Conversely, if one of the equations (5.4) *and* (5.5) *is satisfied by functions u_i, $v_i \in L_p(\Gamma)$, then the other equation is automatically satisfied, and $u_i + iv_i$ equals the interior nontangential limit of its Cauchy integral, almost everywhere on Γ.*

Proof. Suppose f is holomorphic in Ω_i, with $f_i^* \in L_p(\Gamma)$. Let Λ_n be a sequence of simple, closed $C^{1,\alpha}$ curves, and $\gamma_n : \Gamma \to \Lambda_n$ a sequence of invertible $C^{1,\alpha}$ mappings, such that $\Lambda_n \subset \Omega_i$ and

$$\begin{aligned} &\gamma_n(t) \in C_i(t) && \text{for all } t \in \Gamma, \\ &\gamma_n(t) \to t \quad \text{and} \quad \gamma_n'(t) \to 1 && \text{uniformly for } t \in \Gamma. \end{aligned} \tag{5.6}$$

(For example, take a conformal mapping of Ω_i onto the unit disk, and let Λ_n be the inverse image of the circle with radius $1 - 1/n$ and centre zero. The function γ_n can then be defined by requiring the images of t and $\gamma_n(t)$ to both lie on the same radial line. To verify that Λ_n and γ_n have the required properties, use the fact that the conformal mapping is $C^{1,\alpha}$ on $\bar{\Omega}_i$.) If $z \in \Omega_i$, then for all n sufficiently large,

$$f(z) = \frac{1}{2\pi i} \int_{\Lambda_n} \frac{f(t)}{t - z}\, dt = \frac{1}{2\pi i} \int_\Gamma \frac{f(\gamma_n(t))}{\gamma_n(t) - z}\, \gamma_n'(t)\, dt. \tag{5.7}$$

It follows from (5.6) and Theorem 2.2 that $f(\gamma_n(t)) \to f(t)$ for almost all $t \in \Gamma$, and that $|f(\gamma_n(t))| \leqq f_i^*(t)$ for all $t \in \Gamma$. Therefore, we obtain (5.2) by sending $n \to \infty$ in (5.7) and applying the dominated convergence theorem. The first of the two Sokhotski-Plemelj formulae (3.6) now implies $Pf_i = f_i$. Also, using (3.3), it is easy to see that (5.2) is equivalent to (5.3), and that $Pf_i = f_i$ is equivalent to (5.4) and (5.5).

The second part of the theorem follows from the first of the Sokhotski-Plemelj formulae (3.6), once we show that if u_i and v_i satisfy (5.4), then they must also satisfy (5.5), and vice versa.

Suppose $(I-T)u_i=-Rv_i$; then $R(I-T)u_i=-R^2v_i$, and hence $(I+T)Ru_i=(I+T)(I-T)v_i$ by (3.12). Since $I+T$ is invertible on $L_p(\Gamma)$, we conclude that $Ru_i=(I-T)v_i$. A similar argument shows (5.5) implies (5.4). □

It is also possible to derive (5.3)-(5.5) from the boundary integral relations (1.5) and (1.6), using integrations by parts and the Cauchy-Riemann equations (1.2).

Using Theorem 5.1, it is a simple matter to reformulate the interior problem P_i as a 2×2 system of singular integral equations. Denote the characteristic function of Γ_j by χ_j, i.e., let

$$\chi_j(t)=\begin{cases}1 & \text{if } t\in\Gamma_j,\\ 0 & \text{if } t\in\Gamma\backslash\Gamma_j,\end{cases} \tag{5.8}$$

for $j=1$ and 2.

THEOREM 5.2. *Let $g\in L_p(\Gamma)$, and put $g_j=\chi_j g$ for $j=1$ and 2. If $f=u+iv$ is a solution of the interior problem P_i, then the functions*

$$\varphi_1=\chi_1 v_i, \qquad \varphi_2=\chi_2 u_i \tag{5.9}$$

satisfy

$$\begin{bmatrix} R & I-T\\ \chi_2 & \chi_1\end{bmatrix}\begin{bmatrix}\varphi_1\\ \varphi_2\end{bmatrix}=\begin{bmatrix} I-T & -R\\ \chi_2 & \chi_1\end{bmatrix}\begin{bmatrix} g_1\\ g_2\end{bmatrix} \tag{5.10}$$

and

$$\begin{bmatrix} I-T & R\\ \chi_2 & \chi_1\end{bmatrix}\begin{bmatrix}\varphi_1\\ \varphi_2\end{bmatrix}=\begin{bmatrix} R & -(I-T)\\ \chi_2 & \chi_1\end{bmatrix}\begin{bmatrix} g_1\\ g_2\end{bmatrix}. \tag{5.11}$$

Conversely, if φ_1, $\varphi_2\in L_p(\Gamma)$ satisfy either (5.10) *or* (5.11), *and if $f_i=u_i+iv_i$ where*

$$u_i=g_1+\varphi_2, \qquad v_i=\varphi_1+g_2, \tag{5.12}$$

then equation (5.2) *defines a solution f of P_i.*

Proof. Suppose $f=u+iv$ is a solution of P_i. The definitions (5.9) imply (5.12), and when the latter are substituted into (5.4), we find

$$R\varphi_1+(I-T)\varphi_2=-(I-T)g_1-Rg_2. \tag{5.13}$$

Combining this with the equation

$$\chi_2\varphi_1+\chi_1\varphi_2=0=\chi_2 g_1+\chi_1 g_2,$$

we obtain the system (5.10). Similarly, (5.11) follows from (5.5).

Conversely, if φ_1 and φ_2 satisfy either of the systems (5.10) or (5.11), then the functions u_i and v_i satisfy (5.4) and (5.5), and, in addition, $u_i|_{\Gamma_1}=g|_{\Gamma_1}$ and $v_i|_{\Gamma_2}=g|_{\Gamma_2}$. The second part of Theorem 5.1 now shows that the Cauchy integral of u_i+iv_i is a solution of P_i. □

We will now show that the number of solutions in $L_p^2(\Gamma,\rho)$ of (5.10) and (5.11) is determined by the values of

$$\delta_j=\delta(t_{(j)})=2\pi(\beta_j+1/p), \qquad j=1,2. \tag{5.14}$$

In view of assumption (2.1), we have $0<\delta_j<2\pi$ for $j=1$ and 2, and it turns out that three cases need to be distinguished:

1. $0<\delta_j<\pi$ for $j=1$ and 2;
2. Either $0<\delta_1<\pi$ and $\pi<\delta_2<2\pi$, or $\pi<\delta_1<2\pi$ and $0<\delta_2<\pi$;
3. $\pi<\delta_j<2\pi$ for $j=1$ and 2.

Note if $\beta_1=\beta_2=0$, i.e., if $L_p(\Gamma,\rho)=L_p(\Gamma)$, then case 1 occurs when $2<p<\infty$, case 2 is impossible, and case 3 occurs when $1<p<2$.

THEOREM 5.3. *Let A denote any one of the four 2×2 operators appearing in* (5.10) *and* (5.11). *The linear mapping*

$$A: L_p^2(\Gamma,\rho)\to L_p^2(\Gamma,\rho)$$

is bounded, and

$$\ker A^t=\operatorname{span}\{[\psi,0]^t\}, \tag{5.15}$$

where the function ψ is as in Theorem 3.4. *A necessary and sufficient condition for A to be a Fredholm operator is that $\delta_j\neq\pi$ for $j=1$ and 2. When this happens,*

$$\operatorname{ind} A=\begin{cases}-1 \text{ in case } 1,\\ 0 \text{ in case } 2,\\ 1 \text{ in case } 3\end{cases}$$

and, consequently,

$$n(A)=\begin{cases}0 \text{ in case } 1,\\ 1 \text{ in case } 2,\\ 2 \text{ in case } 3.\end{cases} \tag{5.16}$$

Proof. Only $A=\left[\begin{smallmatrix}R & I-T\\ \chi_2 & \chi_1\end{smallmatrix}\right]$ will be discussed—the other three operators can be treated in a similar fashion.

Suppose $[\varphi_1,\varphi_2]^t\in\ker A^t$, i.e.,

$$R^t\varphi_1+\chi_2\varphi_2=0, \tag{5.17}$$

$$(I-T^t)\varphi_1+\chi_1\varphi_2=0, \tag{5.18}$$

and put $u=V\varphi_1+\lambda$, where λ is a constant which will be determined later. Together with Theorem 3.2, (5.17) and (5.18) imply

$$(\partial u/\partial\tau)_i|_{\Gamma_1}=\tfrac{1}{2}R^t\varphi_1|_{\Gamma_1}=0, \tag{5.19}$$

$$(\partial u/\partial\nu)_i|_{\Gamma_2}=-\tfrac{1}{2}(I-T^t)\varphi_1|_{\Gamma_2}=0. \tag{5.20}$$

It follows from (5.19) that u_i is constant on Γ_1 so, after making an appropriate choice of λ, we can assume $u_i|_{\Gamma_1}=0$. Thus, (5.20) and (3.7) now imply u is constant on Ω_i, hence $R^t\varphi_1=(I-T^t)\varphi_1=0$ on all of Γ. This shows $\varphi_1\in\ker R^t=\ker(I-T^t)=\operatorname{span}\{\psi\}$ and $\varphi_2=\chi_1\varphi_2+\chi_2\varphi_2=0$, thereby establishing (5.15).

To prove the remaining assertions, we use (3.3) and (4.1) to write A in the form

$$A=a\mathbf{P}+b\mathbf{Q}+\mathbf{K},$$

where

$$a=\begin{bmatrix}-i & 1\\ \chi_2 & \chi_1\end{bmatrix},\quad b=\begin{bmatrix}i & 1\\ \chi_2 & \chi_1\end{bmatrix},\quad \mathbf{K}=\begin{bmatrix}iT & -T\\ 0 & 0\end{bmatrix}.$$

The operator $\mathbf{K}$ is compact on $L_p^2(\Gamma,\rho)$, and therefore has no effect on the Fredholm property and the index of A. (Note that this is no longer true if Γ is permitted to have

corners; cf. Costabel [2].) An elementary calculation shows

$$b^{-1}a = \begin{bmatrix} \chi_2 - \chi_1 & 0 \\ -2i\chi_2 & 1 \end{bmatrix} = \begin{cases} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} & \text{on } \Gamma_1, \\ \begin{bmatrix} 1 & 0 \\ -2i & 1 \end{bmatrix} & \text{on } \Gamma_2, \end{cases}$$

and

$$\det (b^{-1}a)_{p,\rho}(t,\mu) = \begin{cases} -1 & \text{if } t \in \Gamma_1 \\ 2c_{\delta_1}(\mu) - 1 & \text{if } t = t_{(1)} \\ 1 & \text{if } t \in \Gamma_2 \\ 1 - 2c_{\delta_2}(\mu) & \text{if } t = t_{(2)}. \end{cases} \tag{5.21}$$

Bearing in mind the description of the curve $\mu \mapsto c_\delta(\mu)$ given in § 4, it is not difficult to verify that

$$\operatorname{ind}_{p,\rho}(b^{-1}a) = \begin{cases} 1 & \text{in case 1,} \\ 0 & \text{in case 2,} \\ -1 & \text{in case 3,} \end{cases}$$

and that $b^{-1}a$ fails to be p, ρ-regular precisely when $\delta_1 = \pi$ or $\delta_2 = \pi$. The second part of the theorem now follows from Theorem 4.1, noting that $n(A) = n(A^t) + \operatorname{ind} A$ by (2.5) and (2.6). □

The final theorem for this section establishes the solvability of (5.10) and (5.11), and hence of the interior problem P_i. An argument similar to the proof of the second part of Theorem 5.1 shows that every solution of (5.10) is also a solution of (5.11), and vice versa, and therefore it suffices to discuss only the first of these equations.

THEOREM 5.4. *Assume $\delta_j \neq \pi$ for $j = 1$ and 2, and consider* (5.10). *For every right-hand side $[g_1, g_2]^t \in L_p^2(\Gamma, \rho)$, there exists a solution $[\varphi_1, \varphi_2]^t \in L_p^2(\Gamma, \rho)$, and if $g_1|_{\Gamma_2} = 0$ and $g_2|_{\Gamma_1} = 0$, then likewise $\varphi_1|_{\Gamma_2} = 0$ and $\varphi_2|_{\Gamma_1} = 0$. The homogeneous equation, for which $g_1 = g_2 = 0$, possesses a solution space of dimension* 0, 1, *and* 2 *in cases* 1, 2, *and* 3, *respectively. Moreover, in case* 1, *the linear operator $[\varphi_1, \varphi_2]^t \mapsto [g_1, g_2]^t$ is a continuous isomorphism of $L_p^2(\Gamma, \rho)$ onto itself.*

Proof. Consider $A = \begin{bmatrix} R & I - T \\ \chi_2 & \chi_1 \end{bmatrix}$ and $B = \begin{bmatrix} -(I-T) & -R \\ \chi_2 & \chi_1 \end{bmatrix}$ as operators on $L_p^2(\Gamma, \rho)$. Recall that if the subspace $\operatorname{im} A$ is closed, then it consists precisely of those elements of $L_p^2(\Gamma, \rho)$ which are orthogonal to $\ker A^t$. From Theorem 5.3, we know that $\operatorname{im} A$ and $\operatorname{im} B$ are both closed, and that $\ker A^t = \operatorname{span}\{[\psi, 0]^t\} = \ker B^t$; therefore $\operatorname{im} A = \operatorname{im} B$. The assertions of the theorem now follow from (5.16). □

Before going on to discuss the exterior problem, we will rewrite (5.10) in an alternative form, and give an example in which the nontrivial solutions of the homogeneous problem can be found explicitly.

For $j = 1$ and 2, define $L_p(\Gamma_j, \rho)$ in the obvious way, by setting

$$\|f\|_{L_p(\Gamma_j,\rho)} = \left(\int_{\Gamma_j} |\rho(t) f(t)|^p |dt| \right)^{1/p},$$

where, once again, ρ is given by (5.1). For $j, l \in \{1, 2\}$, we define operators $R_{jl}, T_{j,l} : L_p(\Gamma_{l,\rho}) \to L_p(\Gamma_j, \rho)$ by

$$R_{jl}\varphi_l(z) = \frac{1}{\pi} \int_{\Gamma_l} \frac{\tau(t) \cdot (z - t)}{|z - t|^2} \varphi_l(t) \, |dt|,$$

$$T_{jl}\varphi_l(z) = \frac{1}{\pi} \int_{\Gamma_l} \frac{\nu(t) \cdot (z - t)}{|z - t|^2} \varphi_l(t) \, |dt|,$$

where $z \in \Gamma_j$, and put

$$\varphi_1 = v_i|_{\Gamma_1}, \quad \varphi_2 = u_i|_{\Gamma_2}, \quad g_1 = g|_{\Gamma_1}, \quad g_2 = g|_{\Gamma_2}.$$

Using (5.4), the interior problem P_i can now be reformulated as a 2×2 system of integral equations on $\Gamma_1 \times \Gamma_2$, namely

$$\begin{bmatrix} R_{11} & -T_{12} \\ R_{21} & I - T_{22} \end{bmatrix} \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix} = \begin{bmatrix} -(I - T_{11}) & -R_{12} \\ T_{21} & -R_{22} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}. \tag{5.22}$$

If we agree to identify any function defined on Γ_j with its extension by zero to Γ, then every solution $[\varphi_1, \varphi_2]^t$ of this system is also a solution of (5.10), and vice versa, assuming $g_1|_{\Gamma_2} = 0$ and $g_2|_{\Gamma_1} = 0$. Therefore, Theorem 5.4 shows that, for example, in case 1, the system (5.22) determines a continuous isomorphism $[\varphi_1, \varphi_2]^t \mapsto [g_1, g_2]^t$ of $L_p(\Gamma_1, \rho) \times L_p(\Gamma_2, \rho)$ onto itself. Arguably, (5.22) is more natural than (5.10), even though the latter can be analysed directly using Theorem 4.1.

EXAMPLE 5.5. Let $\Omega_i = \{z \in \mathbb{C}: |z| < 1\}$, and put $t_{(1)} = 1$ and $t_{(2)} = t_{(0)} = -1$. Thus, Γ_1 is the lower half of the unit circle, and Γ_2 is the upper half. The Möbius transformation $w = i(1+z)/(1-z)$ defines a conformal mapping $z \mapsto w$ of the unit disk Ω_i onto the upper half-plane $\{w \in \mathbb{C}: \Im w > 0\}$. Under this transformation, Γ_1 is mapped onto the positive real axis, and Γ_2 is mapped onto the negative real axis, with $t_{(1)} \mapsto \infty$ and $t_{(2)} \mapsto 0$. Write $w = r e^{i\theta}$; then the function

$$w \mapsto i w^{m+1/2} = r^{m+1/2}[-\sin(m+\tfrac{1}{2})\theta + i \cos(m+\tfrac{1}{2})\theta]$$

is holomorphic for $\Im w > 0$, has zero real part when $\theta = 0$ and has zero imaginary part when $\theta = \pi$. Therefore, if the z-plane is cut along the real axis from $-\infty$ to -1, and from $+1$ to $+\infty$, then any branch of the function

$$f(z) = i[i(1+z)/(1-z)]^{m+1/2}$$

is holomorphic in Ω_i and satisfies the homogeneous, mixed boundary conditions

$$u_i|_{\Gamma_1} = 0, \qquad v_i|_{\Gamma_2} = 0,$$

where, as usual, $u = \Re f$ and $v = \Im f$. If $1 < p < 2$ and m is either zero or -1, then $f_i^* \in L_p(\Gamma)$ and hence f is a nontrivial solution of the homogeneous interior problem. More generally, $f_i \in L_p(\Gamma, \rho)$ either when $\pi < \delta_1 < 2\pi$ and $m = 0$, or when $\pi < \delta_2 < 2\pi$ and $m = -1$.

6. The exterior problem. We will say f is holomorphic in Ω_e, if, in addition to being holomorphic in $\Omega_e \setminus \{\infty\}$ in the usual sense, it is also bounded at infinity. In such a case, the function $z \mapsto f(1/z)$ has a removable singularity at $z = 0$, and hence $f(\infty) = \lim_{z\to\infty} f(z)$ exists (cf. the discussion immediately following Theorem 2.2). With this convention, a precise statement of the exterior problem is as follows:

P_e Given $g \in L_p(\Gamma)$, find a function $f = u + iv$ which is holomorphic in Ω_e, has its (exterior) nontangential maximal function $f_e^* \in L_p(\Gamma)$, and satisfies the mixed boundary conditions

$$u_e|_{\Gamma_1} = g|_{\Gamma_1}, \qquad v_e|_{\Gamma_2} = g|_{\Gamma_2}.$$

As with the interior problem, it will be seen that P_e has a unique solution when $2 < p < \infty$. The analysis of P_e is a little more complicated than that of P_i, owing to the changed form of the Cauchy integral formula.

THEOREM 6.1. *If f is holomorphic in Ω_e, and if $f_e^* \in L_p(\Gamma)$, then*

$$f(z) = f(\infty) - \frac{1}{2\pi i} \int_\Gamma \frac{f_e(t)}{t - z}\, dt, \qquad z \in \Omega_e, \tag{6.1}$$

and $Qf_e = f_e - f(\infty)$. *Moreover,*

$$f(\infty) = \langle \psi, f_e \rangle, \tag{6.2}$$

where ψ *is as in Theorem* 3.4.

Proof. Suppose f satisfies the hypotheses of the theorem, and let $\Omega_r = \{z \in \Omega_e : |z| < r\}$ and $\Gamma_r = \{z \in \mathbb{C} : |z| = r\}$. Given $z \in \Omega_e$, an argument similar to the proof of Theorem 5.1 shows that, for all r sufficiently large,

$$f(z) = \frac{1}{2\pi i} \int_{\Gamma_r} \frac{f(t)}{t-z}\, dt - \frac{1}{2\pi i} \int_{\Gamma} \frac{f_e(t)}{t-z}\, dt, \tag{6.3}$$

and since $f(t) = f(\infty) + O(t^{-1})$ as $t \to \infty$, it follows that

$$\frac{1}{2\pi i} \int_{\Gamma_r} \frac{f(t)}{t-z}\, dt = f(\infty) + O(r^{-1}).$$

Therefore, by sending $r \to \infty$ in (6.3), we obtain (6.1).

The second of the two Sokhotski-Plemelj formulae (3.6) now implies $Qf_e = f_e - f(\infty)$, and since $Q^t\psi = \frac{1}{2}[(I - T^t)\psi - iR^t\psi] = 0$, we see

$$\langle \psi, f_e - f(\infty) \rangle = \langle \psi, Qf_e \rangle = \langle Q^t\psi, f_e \rangle = 0.$$

This establishes (6.2), because $\langle \psi, f(\infty) \rangle = f(\infty)$. □

Introduce the linear functional

$$L(u) = 2\langle \psi, u \rangle = 2 \int_{\Gamma} \psi(t) u(t) |dt|;$$

then, when $f = u + iv$, the equation $Qf_e = f_e - f(\infty)$ is equivalent to the pair of equations

$$(I + T)u_e - L(u_e) = Rv_e, \tag{6.4}$$

$$(I + T)v_e - L(v_e) = -Ru_e. \tag{6.5}$$

Moreover, an argument similar to the last part of the proof of Theorem 5.1 shows that every pair of functions u_e, $v_e \in L_p(\Gamma)$ that satisfies (6.4) automatically satisfies (6.5), and vice versa. Consequently, we arrive at the following reformulation of the exterior problem, where, as before, χ_j is defined by (5.8).

THEOREM 6.2. *Let* $g \in L_p(\Gamma)$, *and put* $g_j = \chi_j g$ *for* $j = 1$ *and* 2. *If* $f = u + iv$ *is a solution of the exterior problem* P_e, *then the functions*

$$\varphi_1 = \chi_1 v_e, \qquad \varphi_2 = \chi_2 u_e$$

satisfy

$$\begin{bmatrix} R & L - (I+T) \\ \chi_2 & \chi_1 \end{bmatrix} \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix} = \begin{bmatrix} (I+T) - L & -R \\ \chi_2 & \chi_1 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \tag{6.6}$$

and

$$\begin{bmatrix} (I+T) - L & R \\ \chi_2 & \chi_1 \end{bmatrix} \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix} = \begin{bmatrix} -R & L - (I+T) \\ \chi_2 & \chi_1 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}. \tag{6.7}$$

Conversely, if functions $\varphi_1, \varphi_2 \in L_p(\Gamma)$ *satisfy either* (6.6) *or* (6.7), *and if* $f_e = u_e + iv_e$, *where*

$$u_e = g_1 + \varphi_2, \qquad v_e = \varphi_1 + g_2,$$

then the function f, *defined by* (6.1) *and* (6.2), *is a solution of* P_e.

As with the interior problem, the number of solutions in $L_p^2(\Gamma, \rho)$ of (6.6) and (6.7) depends on the values of δ_1 and δ_2 in (5.14).

THEOREM 6.3. *All of the assertions made in Theorem* 5.3 *about the* 2×2 *operators in* (5.10) *and* (5.11) *remain true for the* 2×2 *operators in* (6.6) *and* (6.7).

Proof. Only $A=\left[\begin{smallmatrix} R & L-(I+T) \\ \chi_2 & \chi_1 \end{smallmatrix}\right]$ will be discussed—the other three operators can be treated in a similar fashion.

Suppose $[\varphi_1, \varphi_2]^t \in \ker A^t$, i.e.,

$$R^t\varphi_1+\chi_2\varphi_2=0, \tag{6.8}$$

$$2\langle\varphi_1, 1\rangle\psi-(I+T^t)\varphi_1+\chi_1\varphi_2=0, \tag{6.9}$$

and put $u=V(\varphi_1-\langle\varphi_1, 1\rangle\psi)+\lambda$, where λ is a constant whose value will be chosen later. The results of Theorem 3.4 imply

$$\begin{aligned}(\partial u/\partial\tau)_e&=\tfrac{1}{2}R^t\varphi_1,\\ (\partial u/\partial\nu)_e&=\tfrac{1}{2}(I+T^t)\varphi_1-\langle\varphi_1, 1\rangle\psi,\end{aligned} \tag{6.10}$$

and therefore, by (6.8) and (6.9),

$$(\partial u/\partial\tau)_e|_{\Gamma_1}=0, \qquad (\partial u/\partial\nu)_e|_{\Gamma_2}=0. \tag{6.11}$$

The first of these equations implies u is constant on Γ_1, so an appropriate choice of λ gives $u_e|_{\Gamma_1}=0$. Combining this fact with the second equation in (6.11), and noting that $\int_\Gamma(\varphi_1-\langle\varphi_1, 1\rangle\psi)|dt|=0$, we see from Lemma 3.3 that u must be constant on Ω_e; in fact, $u|_{\Omega_e}=u(\infty)=\lambda$. Equations (6.10) now imply $\varphi_1\in\ker R^t=\operatorname{span}\{\psi\}$, and hence it follows from (6.8) and (6.9) that $\chi_2\varphi_2=0=\chi_1\varphi_2$, so $\varphi_2=0$. This completes the proof of (5.15).

Next, write A in the form

$$A=a\mathbf{P}+b\mathbf{Q}+\mathbf{K},$$

putting, in view of (3.3) and (4.1),

$$a=\begin{bmatrix}-i & -1\\ \chi_2 & \chi_1\end{bmatrix}, \quad b=\begin{bmatrix} i & -1\\ \chi_2 & \chi_1\end{bmatrix}, \quad \mathbf{K}=\begin{bmatrix} iT & L-T\\ 0 & 0\end{bmatrix}.$$

We find that

$$b^{-1}a=\begin{bmatrix}\chi_2-\chi_1 & 0\\ 2i\chi_2 & 1\end{bmatrix}=\begin{cases}\begin{bmatrix}-1 & 0\\ 0 & 1\end{bmatrix} & \text{on } \Gamma_1\\ \begin{bmatrix}1 & 0\\ 2i & 1\end{bmatrix} & \text{on } \Gamma_2,\end{cases}$$

and $\det(b^{-1}a)_{p,\rho}(t,\mu)$ is the same as in (5.21). Thus, the remainder of the proof simply repeats that of Theorem 5.3. □

We claim that every solution of (6.6) is also a solution of (6.7). Indeed, if both sides of the equation

$$R\varphi_1+L(\varphi_2)-(I+T)\varphi_2=(I+T)g_1-L(g_1)-Rg_2$$

are multiplied on the left by $-R$, then the results of Theorem 3.4 imply

$$(I-T)[(I+T)\varphi_1+R\varphi_2]=(I-T)[-Rg_1-(I+T)g_2],$$

and so there is a constant λ such that

$$(I+T)\varphi_1+R\varphi_2=-Rg_1-(I+T)g_2+\lambda.$$

Applying the functional L to both sides of this equation, and using known properties of the function ψ, we find $\lambda=L(\varphi_1+g_2)$, as required. A similar argument shows that every solution of (6.7) is also a solution of (6.6), so it suffices to discuss only the latter equation, and the reader may easily verify the following theorem.

THEOREM 6.4. *All of the assertions made in Theorem* 5.4 *about* (5.10) *remain true for* (6.6).

Just as with the interior problem, it is possible to reformulate the exterior problem as a 2×2 system of integral equations on $\Gamma_1\times\Gamma_2$, instead of using (6.6) or (6.7), but we will not bother to do this. Moreover, when Γ is the unit circle, the nontrivial solutions of the homogeneous exterior problem can be exhibited explicitly as follows.

EXAMPLE 6.5. Let Γ_1 and Γ_2 be the same as in Example 5.5, and note that $\Omega_e=\{z\in\mathbb{C}\colon |z|>1\}\cup\{\infty\}$. The conformal mapping defined by $z\mapsto w=i(1+z)/(1-z)$ takes Ω_e onto the lower half-plane $\{w\in\mathbb{C}\colon \Im w<0\}$, and, once again, we put

$$f(z)=i[i(1+z)/(1-z)]^{m+1/2},$$

where m is any integer. This time, however, the plane is cut along the real axis between -1 and $+1$. Then any branch of f is holomorphic in Ω_e with $f(\infty)=i(-i)^{m+1/2}$, and satisfies

$$u_e|_{\Gamma_1}=0, \qquad v_e|_{\Gamma_2}=0,$$

where, as usual, $u=\Re f$ and $v=\Im f$. Thus, for $1<p<2$ and $m=0$ or -1, the function f is a nontrivial solution of the homogeneous exterior problem. Also, $f_e\in L_p(\Gamma,\rho)$ when $\pi<\delta_1<2\pi$ and $m=0$, or when $\pi<\delta_2<2\pi$ and $m=-1$.

REFERENCES

[1] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[2] M. COSTABEL, *Singular integral operators on curves with corners*, Integral Equations Operator Theory, 3 (1980), pp. 323-349.

[3] M. COSTABEL AND E. STEPHAN, *Boundary integral equations for mixed boundary value problems in polygonal domains and Galerkin approximation*, in Mathematical Models and Methods in Mechanics, Banach Center Publications, Vol. 15, PWN-Polish Scientific Publishers, Warsaw, Poland, 1985, pp. 175-251.

[4] B. E. J. DAHLBERG, *On the Poisson integral for Lipschitz and C^1 domains*, Studia Math., 66 (1979), pp. 13-24.

[5] E. B. FABES, M. JODEIT, JR., AND N. M. RIVIÈRE, *Potential techniques for boundary value problems on C^1 domains*, Acta Math., 141 (1978), pp. 165-186.

[6] I. GOHBERG AND N. KRUPNIK, *Einführung in die Theorie der eindimensionalen singulären Integraloperatoren*, Birkhäuser, Basel, Switzerland, 1979.

[7] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[8] M. A. JASWON AND G. T. SYMM, *Integral Equation Methods in Potential Theory and Elastostatics*, Academic Press, New York, 1977.

[9] C. E. KENIG, *Elliptic boundary value problems on Lipschitz domains*, in Beijing Lectures in Harmonic Analysis, Ann. Math. Stud., 112, E. M. Stein, ed., Princeton University Press, Princeton, NJ, 1986, pp. 131-183.

[10] J. KRÁL, *Integral Operators in Potential Theory*, Lecture Notes in Math. 823, Springer-Verlag, Berlin, New York, 1982.

[11] S. G. KREIN, *Linear Equations in Banach Spaces*, Birkhäuser, Boston, 1982.

[12] S. G. MIKHLIN, *Mathematical Physics, An Advanced Course*, North-Holland, Amsterdam, 1970.

[13] S. G. MIKHLIN AND S. PRÖSSDORF, *Singular Integral Operators*, Springer-Verlag, Berlin, New York, 1986.

[14] N. I. MUSKHELISHVILI, *Singular Integral Equations*, Noordhoff, Groningen, the Netherlands, 1953.

[15] V. I. SMIRNOV, *A Course of Higher Mathematics*, Vol. IV, Pergamon, Oxford, 1964.

[16] W. L. WENDLAND, E. STEPHAN, AND G. C. HSIAO, *On the integral equation method for the plane mixed boundary value problem of potential theory*, Math. Methods Appl. Sci., 1 (1979), pp. 265-321.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 935-953, July 1990
© 1990 Society for Industrial and Applied Mathematics
007

MULTIPLE COUPLING IN CHAINS OF OSCILLATORS*

N. KOPELL†, W. ZHANG‡, AND G. B. ERMENTROUT§

Abstract. Chains of oscillators with coupling to more neighbors than the nearest ones are considered. The equations for the phase-locked solutions of an infinite chain of such type may be considered as a one-parameter family of $(2m-1)$st-order discrete dynamical systems, whose independent variable is position along the chain, whose dependent variable is the phase between successive oscillators, and where m is the number of neighbors connected to each side. It is shown that for each value of the parameter in some range, the $(2m-1)$st-order system has a one-dimensional hyperbolic global center manifold. This is done by using the theory of exponential dichotomies to show that the system "shadows" a simple one-dimensional system. The exponential dichotomy is constructed by exploiting an algebraic structure imposed by the geometry of the multiple coupling.

For a finite chain, the dynamical system is constrained by manifolds of boundary conditions. It is shown that for open sets of such conditions, the solution to the equation for phase-locking in long chains stays close to the center manifold except near the boundaries. This is used to show that a multiply coupled system behaves, except near the boundaries, as a modified nearest-neighbor system. The properties of the nearest-neighbor and multiply coupled systems are then compared.

Key words. oscillators, exponential dichotomy, neural networks, central pattern generator, invariant manifold, singular perturbation

AMS(MOS) subject classifications. 34, 58

1. Introduction.

Chains of oscillators with nearest-neighbor coupling have been investigated in [1]-[3]. We showed [2], [3] that the phase-locked solutions of such chains could be approximated, when there is a large number of oscillators, by a discretization of a solution to a singularly perturbed second-order two-point boundary value problem. Thus, over much of the chain, the solution behaves like a solution to a first-order "outer equation"; the particular "outer solution" that is expressed is determined by the boundary conditions for the BVP.

In this paper, we consider coupling that extends beyond *nearest* neighbors to *multiple* neighbors. We will show that the new equations for the phase relationships among the phase-locked oscillators are of order $2m$, where m is the number of neighbors on each side to which any oscillator away from the boundaries of the chain is connected (see Fig. 1.1). Nevertheless, if the boundaries are ignored, it still holds that there are solutions to this $2m$th-order equation that behave like one of a family of solutions to a simple one-dimensional "outer equation." When the boundary effects are taken into

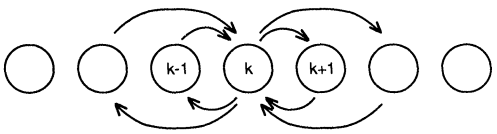


FIG. 1.1. *Schematic diagram of a chain of oscillators, with each oscillator coupled to m neighbors on each side. In the figure* $m=2$, *and only the connections to and from oscillator k are explicitly shown.*

* Received by the editors November 21, 1988; accepted for publication (in revised form) July 28, 1989. This research was partially supported by National Science Foundation grants DMS 8796235 and DMS 537196 and by the Air Force Office of Scientific Research under U.R.I. contract F49620-86-C-0131 to Northeastern University.

† Department of Mathematics, Boston University, Boston, Massachusetts 02215.

‡ Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

§ Department of Mathematics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

account, they again determine which outer solution is chosen. We also compare properties of nearest-neighbor and multiple coupling, and derive some differences that may be important for applications, in particular to the neural network that is the central pattern generator for locomotion in fishlike species [4].

As in [1]–[3], each oscillator of the chain is described by a differential equation of arbitrary dimension, whose only restriction is that it has a stable limit cycle solution. The oscillators may be different from one another, but they must vary in a gradual way along the chain, i.e., if there are $N+1$ oscillators, and if the kth oscillator is described by

$$X_k' = F_k(X_k) \tag{1.1}$$

then $F_{k+1} - F_k = 0(1/N)$.

The coupling is assumed to be pairwise among neighbors whose indices differ by no more than m. (For nearest-neighbor coupling, $m = 1$.) That is, the full equations have the form

$$X_k' = F_k(X_k) + \sum_{j=1}^{m} G_j^+(X_{k+j}, X_k) + \sum_{j=1}^{m} G_j^-(X_{k-j}, X_k). \tag{1.2}$$

The number of coupled neighbors is assumed to be small relative to N. The chain is finite, so $k = 1, \cdots, N+1$; the first m and the last m equations must therefore be modified. This can be done in various ways, as will be described in § 4. Equations (1.2) may be generalized to allow the strength of coupling to vary with position along the chain as in [3], but we will not do so here.

In the absence of the coupling terms, there is an $N+1$-dimensional invariant torus that is the product of the limit cycles of the individual oscillators. The coupling is not assumed to be "weak," but it *is* assumed to be not so strong as to upset the existence of an invariant manifold. If the limit cycles are relatively robust, this is not a restrictive hypothesis. (The size of the allowable perturbation that maintains the existence of the invariant manifold is related to the size of the Floquet exponents and is independent of the number of oscillators. See the Appendix of [2] for further discussion.) With this assumption, (1.2) on the invariant torus takes the form

$$\theta_k' = \omega_k + \sum_{j=1}^{m} h_j^+(\theta_{k+j}, \theta_k) + \sum_{j=1}^{m} h_j^-(\theta_{k-j}, \theta_k). \tag{1.3}$$

Here, θ_k is the phase and ω_k is the frequency of the kth oscillator. h_j^+ and h_j^- are 2π-periodic scalar functions of their arguments. The derivation of (1.3) is similar to that of the special case $m = 1$, which is in [1].

The more restrictive hypothesis to be used in developing the continuum description of the equation is that the phase-locking behavior is not significantly changed if the method of averaging is applied to (1.3). This is automatically valid if the coupling is sufficiently weak [5], and may sometimes fail drastically if it is not [6]. However, weak coupling is not necessary for the application of the method of averaging; as discussed in [7], there are other hypotheses under which the averaged equations differ from the full ones by a small amount that does not affect the existence of phase-locked solutions. The great advantage of the averaged equations is that the coupling terms in (1.3) are replaced by terms that depend only on the *differences* of the phases. The equations then have the form

$$\theta_k' = \omega_k + \sum_{j=1}^{m} H_j^+(\theta_{k+j} - \theta_k) + \sum_{j=1}^{m} H_j^-(\theta_{k-j} - \theta_k) \tag{1.4}$$

where H^+ and H^- are 2π-periodic functions of their arguments [1], [2]. It is (1.4) that we will be investigating. The hypotheses on $H^{\pm}$, which are quite general, are essentially as in [2], [3] and will be spelled out in § 2. We will be particularly interested in phase-locked solutions, i.e., solutions for which θ'_k is independent of k; for equations of the form (1.4), θ'_k is also independent of t. It will be useful to consider equivalent equations, with the variables $\{\theta_k\}$ replaced by $\{\phi_k = \theta_{k+1} - \theta_k\}$. If $\theta'_k \equiv \Omega$, an unknown constant independent of k, (1.4) becomes

$$\Omega = \omega_k + \sum_{j=1}^{m} H_j^+(\phi_{k+j-1} + \cdots + \phi_k) + \sum_{j=1}^{m} H_j^-(-\phi_{k-j} - \cdots - \phi_{k-1}). \tag{1.5}$$

In (1.5), Ω is the frequency of the phase-locked ensemble of oscillators. In general, even if all the oscillators have the same uncoupled frequency, Ω need not equal this common frequency [2], [6] (see § 2).

In addition to the hypotheses on H_j^+ and H_j^-, § 2 contains some frequently used notation and a brief review of the analogous theory for nearest-neighbor coupling. We also define a simple one-parameter family of "outer solutions" $\{\Phi_k\}$ for (1.5), parameterized by Ω, that generalize the one-parameter family of outer solutions for $m = 1$. The main analytic results of the paper are given in § 3. We show that $\{\phi_k\}$ have the property that, for each such outer solution (for a range of Ω), there are solutions to (1.5) that stay arbitrarily close for $1 \leqq k \leqq N$, i.e., that shadow the outer solution. Thus, except near the boundaries, the complicated equation (1.5) is shown to have solutions that behave in a quite simple manner. This is demonstrated by exploiting the special structure of (1.5) to show that the linearization of (1.5) around each outer solution $\{\Phi_k\}$ is hyperbolic. We then make use of a discrete version of the theory of exponential dichotomies [8], [9] which implies the appropriate shadowing lemma.

Section 4 discusses the structure of the solutions to the full equations, with constraints due to finite boundaries. For nearest-neighbor ($m = 1$) coupling, it is shown in [2], [3] that, if the solution to the associated boundary value problem stays within certain limits, there *is* a (unique) solution to (1.5) and associated boundary conditions

$$\Omega = \omega_1 + H^+(\phi_1), \qquad \Omega = \omega_N + H^-(\phi_{N-1}). \tag{1.6}$$

This solution behaves like the outer solution of the continuum boundary value problem except near zero or N, where there is a discrete version of a boundary layer. (See § 2 for a brief review of that theory.) For $m > 1$, there is no longer a natural continuum boundary value problem whose solution is the limit of the solution to (1.5) (with some analogue of (1.6)) as $N \to \infty$. However, the problem for a finite chain of oscillators, multiply coupled, can still be formulated in terms of a dynamical system, with boundary value conditions reminiscent of Sil'nikov type problems [10]. Furthermore, for an open set of choices of boundary conditions, the solution, if it exists, is locally unique and close, over most of the interval $[0, N]$ to one of the outer solutions $\{\Phi_k\}$. For $m > 1$, there is generically a boundary layer on each end. The value of Ω associated with the solution is defined implicitly by the intersection of certain manifolds, and is not computed explicitly, unlike the case $m = 1$ [2], [3]. Nevertheless, approximations to Ω *can* be computed, and we show how. We have not for $m > 1$ ruled out global nonuniqueness for a given set of boundary conditions, e.g., a pair of solutions corresponding to different frequencies Ω.

In § 5 we compare nearest-neighbor coupling and multiple coupling. Using the approximations to Ω mentioned above, we show that one effect of multiple coupling is to reduce the phase differences $\{\phi_k\}$ from what they would be for nearest-neighbor coupling. We note that this is not an effect of merely strengthening the coupling; for

example, if the frequencies ω_k are constant, the phase differences are independent of the strength of coupling for nearest-neighbor coupling [2]. A second effect is that the solution becomes more stable to perturbations of parameters of the problem such as local natural frequencies ω_k. Both of these effects are relevant to the functioning of the locomotion central pattern generator mentioned above.

2. Outer solutions and review of nearest-neighbor theory. In this section, we first introduce some notation and give the equation defining $\{\Phi_k\}$. Before further analysis, we then review some of the theory for chains with nearest-neighbor coupling [2], [3].

Let f_j and g_j, $1 \leqq j \leqq m$, be defined by

$$f_j(\phi)+g_j(\phi)=H_j^+(\phi), \qquad f_j(\phi)-g_j(\phi)=H_j^-(-\phi). \tag{2.1}$$

For each Ω, define $\{\Phi_k\}$ by

$$\Omega=\omega_k+2\sum_j f_j(j\Phi_k)\equiv\omega_k+2F(\Phi_k). \tag{2.2}$$

This solution exists provided that $\{\Phi_k\}$ stay in the region in which $F(\phi)$ is monotone, and hence invertible. Note that in the kth equation, the same Φ_k is used in the argument for each f_j. The $\{\Phi_k\}$ will be the analogue of the "outer solution" for a singularly perturbed equation, as explained later. We will assume the following.

H1. There is a closed interval $I \in S^1$ such that $F'(\phi) \neq 0$ for $\phi \in I$, and the relevant solutions to (2.2) stay in that interval. More strongly, for each j, $f_j'(j\phi) \neq 0$, with the same sign for all j.

H2. For each j, $|f_j'(j\phi)| < g_j'(j\phi)$, so that $H_j^+(j\phi)$ is monotone increasing and $H_j^-(-j\phi)$ is monotone decreasing in I.

Note that if these conditions are valid for some Ω, they hold for an open set of Ω. The techniques of § 3 also require that $|\omega_k-\omega_j|$ remain sufficiently small for all j, k.

As a simple example, consider $H^+(\phi)=H^-(\phi)=A\cos\phi+B\sin\phi$ and $H_j^\pm=\alpha_j H^\pm$, $a_j>0$. (See [1] for oscillators and coupling that give rise to such functions $H^\pm$.) Within a region around $\phi=0$, $H^\pm$ is a monotone increasing function of ϕ. To keep $f'(j\phi)=-A\sin(j\phi)\neq 0$, it is necessary to restrict I to one side of $\phi=0$. This is easy to do for a range of Ω if $A\neq 0$, and for all k, $|\omega_k-\omega_1|$ is not too large.

If $m=1$, (2.2) is

$$\Omega=\omega_k+2f(\Phi_k) \tag{2.3}$$

and (1.5) is

$$\Omega=\omega_k+f(\phi_k)+f(\phi_{k-1})+g(\phi_k)-g(\phi_{k-1}). \tag{2.4}$$

By subtracting the equations of (2.4) pairwise, we can rewrite that set of equations as

$$0=(\omega_{k+1}-\omega_k)+f(\phi_{k+1})-f(\phi_k)+g(\phi_{k+1})-2g(\phi_k)+g(\phi_{k-1}). \tag{2.5}$$

It is shown in [2] that for N large (2.5) formally approaches the continuum equation

$$0=\beta(x)+2f(\phi)_x+\frac{1}{N}g(\phi)_{xx} \tag{2.6}$$

where $0 \leqq x \leqq 1$ and $\beta(x)$ is a continuum limit for $\{\beta(k/N)=\beta_k\equiv(\omega_{k+1}-\omega_k)/N\}$. For N large, (2.6) is a singularly perturbed equation with "outer equation"

$$0=\beta(x)+2f(\phi)_x \tag{2.7}$$

whose integrated form is

$$\Omega=\omega(x)+f(\phi) \tag{2.8}$$

where $\omega(x) \equiv \int_0^x \beta(s)\, ds$ and Ω is a constant of integration. Thus, (2.2) is an extension to multiple coupling of a discrete version of (2.8).

In [2], we showed that the solution to (1.5), (1.6), when it exists, converges as $N \to \infty$ to the solution to (2.5) with the boundary conditions

$$H^-(-\phi) = 0 \quad \text{at } x = 0, \qquad H^+(\phi) = 0 \quad \text{at } x = 1. \tag{2.9}$$

(The convergence is nonuniform near the boundary layer; better convergence may be obtained by using more refined, but more complicated, boundary conditions [3].) It can be seen that (2.9) is the natural limit of conditions (1.6), which, taken together with (1.5), are equivalent to $H^-(-\phi_0) = 0 = H^+(\phi_{N+1})$. The proof mimics, for discrete equations, a proof of the existence of a solution to (2.6), (2.9). In addition to the outer equation (2.3), we define an "inner equation"

$$0 = f(\phi_{k+1}) - f(\phi_k) + g(\phi_{k+1}) - 2g(\phi_k) + g(\phi_{k-1}). \tag{2.10}$$

We show that there are solutions to (1.5) that "shadow" the outer solutions (2.3) and solutions that shadow the solutions to (2.10) for $k \leqq 0(\sqrt{N})$ (or $N - k \leqq 0(\sqrt{N})$). By properly choosing these shadowing solutions and matching them at two successive values of k, we can construct a solution to (1.5), (1.6) with $m = 1$. By the construction, the convergence result is then clear.

The frequency of the locked ensemble can be read off from the outer equation (2.7). For example, if the outer solution satisfies $\phi = \phi_L$ at $x = 0$ (respectively, $\phi = \phi_R$ at $x = 1$) then, as $N \to \infty$, the ensemble frequency tends to the value given by (2.8) i.e., $\Omega = \omega(0) + f(\phi_L)$ (respectively, $\Omega = \omega(1) + f(\phi_R)$).

In [3], the results of [2] for $m = 1$ are extended to deal with cases in which the outer solution $\{\Phi_k\}$ is allowed to be on both sides of the "turning point" of (2.6), i.e., $f' = F'$ is not required to be bounded away from zero on the outer solution. This leads to the possibility of "phase transition"-like behavior in the solutions in which, in the limit as $N \to \infty$, the solution to the boundary value problem changes discontinuously as some parameters are changed continuously. By requiring that $F'(\phi) \neq 0$, we are avoiding that extra set of complications.

3. The hyperbolic structure of the outer solutions and the shadowing lemma. In this section, we generalize the part of the above proof for $m = 1$ dealing with solutions to (1.5) that shadow outer solutions. We first show that if (1.5) is linearized around an outer solution the resulting linear system with nonconstant coefficients is hyperbolic at each k, with the dimensions of the splitting independent of k. This can be done in the absence of further information about $\{f_i\}$ and $\{g_i\}$ by exploiting the structure of the equations to show that the characteristic polynomial of the linearization at each k has a special form that allows us to compute the number of eigenvalues less than and greater than 1 in absolute value.

Using Taylor series and the definition (2.2) of $\{\Phi_k\}$, (1.5) can be rewritten as

$$\begin{aligned} 0 = &\sum_{j=1}^{m} (f_j + g_j)'(j\Phi_k)(\phi_k + \cdots + \phi_{k+j-1} - j\Phi_k) \\ &+ \sum_{j=1}^{m} (f_j - g_j)'(j\Phi_k)(\phi_{k-1} + \cdots + \phi_{k-j} - j\Phi_k) \\ &+ \sum_{ij} 0((\phi_i - \Phi_k)(\phi_j - \Phi_k)) \end{aligned} \tag{3.1}$$

where the last term denotes all the nonlinear parts of the Taylor expansion. We turn

(3.1) into a first-order system using the $2m-1$ coordinates

$$\begin{aligned} &\eta_k = \phi_k - \Phi_k, && z_{k,1} = \eta_{k+1}, \\ &x_{k,1} = \eta_{k-1}, && \vdots \\ &\vdots && z_{k,m-1} = \eta_{k+m-1}. \\ &x_{k,m-1} = \eta_{k-m+1}, \end{aligned}$$

In these coordinates, (3.1) is a first-order system. The first $2m-2$ equations come from the definitions of the new variables and are given simply by

$$\begin{aligned} &x_{k+1,i} = x_{k,i-1}, && i>1, \\ &x_{k+1,1} = \eta_k, \\ &\eta_{k+1} = z_{k,1}, \\ &z_{k+1,i} = z_{k,i+1}, && i<m-1. \end{aligned}$$

To compute the last equation, we use the $(k+1)$st equation of (3.1), and note that for $j>m$,

$$\begin{aligned} (\phi_{k+1}+\cdots+\phi_{k+j}-j\Phi_{k+1}) &= \sum_{i=1}^{j} (\phi_{k+i}-\Phi_{k+1}) \\ &= z_{k,1}+\cdots+z_{k,j}+\sum_{i=1}^{j} (\Phi_{k+i}-\Phi_{k+1}). \end{aligned}$$

For $j=m$,

$$(\phi_{k+1}+\cdots+\phi_{k+m}-j\Phi_{k+1}) = z_{k,1}+\cdots+z_{k,m-1}+z_{k+1,m-1}+\sum_{i=1}^{m} (\Phi_{k+i}-\Phi_{k+1}).$$

Similarly,

$$(\phi_k+\cdots+\phi_{k-j+1}-j\Phi_{k+1}) = \eta_k+x_{k,1}+\cdots+x_{k,j-1}+\sum_{i=0}^{j-1} (\Phi_{k-i}-\Phi_{k+1}).$$

We note that the sums involving terms of the form $\Phi_{k\pm i}-\Phi_{k+1}$ are all $0(1/N)$, since j is assumed to be small relative to N.

We now let V_j (respectively, W_j) denote $(f_j+g_j)'(j\Phi_{k+1})$ (respectively, $(f_j-g_j)'(j\Phi_{k+1})$). Then the $(k+1)$st equation of (3.1) may be rewritten as

$$\begin{aligned} 0 = {} & W_m x_{k,m-1}+(W_m+W_{m-1})x_{k,m-2}+\cdots+\left(\sum_{j=1}^{m} W_j\right)\eta_k \\ & +\left(\sum_{j=1}^{m} V_j\right) z_{k,1}+\cdots+(V_m+V_{m-1})z_{k,m-1}+V_m z_{k+1,m-1} \\ & +0(1/N)+\mathcal{N}(x_{k,i},\, \eta_k,\, z_{k,i}) \end{aligned} \tag{3.2}$$

where $\mathcal{N}$ is at least quadratic in its variables. Thus, the linearization L_{k+1} of (1.5) at Φ_{k+1} is a $(2m-1)\times(2m-1)$ matrix with 1's to the right of the main diagonal and zeros elsewhere in the first $2m-2$ rows. The last row is

$$\frac{1}{V_m}\left(W_m,\, W_m+W_{m-1},\, \cdots,\, \sum_{j=1}^{m} W_j,\, \sum_{j=1}^{m} V_j,\, \cdots,\, V_m+V_{m-1}\right).$$

From the simple form of this matrix, it is easy to see the following lemma.

LEMMA 3.1. *Let* $C(\lambda) \equiv V_m \det(\lambda I - L_{k+1})$. *Then*

$$C(\lambda) = V_m\lambda^{2m-1} + (V_m + V_{m-1})\lambda^{2m-2} + \cdots + \left(\sum_{j=1}^{m} V_j\right)\lambda^m + \left(\sum_{j=1}^{m} W_j\right)\lambda^{m-1} + \cdots + (W_m + W_{m-1})\lambda + W_m. \tag{3.3}$$

□

Recall that, by hypothesis, $V_j > 0$ and $W_j < 0$ for all j and all Φ_k. We can use this to compute the number of eigenvalues having absolute value less than 1 (respectively, greater than 1). We will show the following lemma.

LEMMA 3.2. $C(\lambda)$ *has no zeros on* $|\lambda| = 1$. *If* $2F' \equiv \sum_{j=1}^{m} (jW_j + jV_j) > 0$ (*respectively, less than zero*), *there are m eigenvalues inside* $|\lambda| = 1$ *and* $m-1$ *eigenvalues outside* $|\lambda| = 1$ (*respectively,* $m-1$ *inside and m outside* $|\lambda| = 1$).

Proof. The main tool is the theorem of Pellet [11], [12] which can be proved by an argument using Rouché's theorem. The case we use can be stated as follows.

THEOREM [12]. *Suppose a polynomial*

$$P(\lambda) = a_0 + a_1\lambda + \cdots + a_p\lambda^p + \cdots + a_n\lambda^n$$

with real coefficients satisfies $a_i > 0$, $i \neq p$, $a_p < 0$, *and* $P(\lambda)$ *has two positive roots at r and R, with* $r < R$. *Then P has exactly p zeros in or on the circle* $|\lambda| \leqq r$ *and no zeros in the annulus* $r < \lambda < R$.

First, suppose $\sum_{j=1}^{m} (jW_j + jV_j) > 0$. To use Pellet's theorem, we let

$$\begin{aligned} P(\lambda) &= (\lambda - 1)C(\lambda) \\ &= V_m\lambda^{2m} + V_{m-1}\lambda^{2m-1} + \cdots + V_1\lambda^{m+1} + \left[\sum_{j=1}^{m} (W_j - V_j)\right]\lambda^m - W_1\lambda^{m-1} \cdots - W_m. \end{aligned}$$

Since $V_j > 0$ and W_j is negative, $P(\lambda)$ satisfies $a_i > 0$, $i \neq m$, $a_m < 0$. Also $P(1) = 0$. To see that there is another root $r < 1 \equiv R$, note that $P(0) > 0$. By hypothesis, $C(1) = \sum_{j=1}^{m} (jW_j + jV_j) > 0$, so $C(\lambda) > 0$ in a neighborhood of $\lambda = 1$. Hence $P(\lambda) < 0$ for real $\lambda < 1$ and $P(\lambda) > 0$ for $\lambda > 1$. This ensures the existence of the root $r < 1$ of $P(\lambda)$. By Pellet's theorem, $P(\lambda)$ has exactly m roots satisfying $|\lambda| \leqq r < 1$ and no more inside $|\lambda| = 1$. Thus $C(\lambda)$ has exactly m roots inside $|\lambda| = 1$.

We now show that there are $m-1$ roots outside $|\lambda| = 1$. Let

$$\bar{C}(\lambda) = -W_m\lambda^{2m-1} - (W_m + W_{m-1})\lambda^{2m-2} - \cdots - \left(\sum_{j=1}^{m} W_j\right)\lambda^m - \left(\sum_{j=1}^{m} V_j\right)\lambda^{m-1} - \cdots - V_m.$$

The roots of $\bar{C}(\lambda)$ are the inverses of the roots of $C(\lambda)$, so it suffices to show that $\bar{C}(\lambda)$ has $m-1$ roots inside $|\lambda| = 1$. Let $\gamma \in \mathbf{R}$ satisfy

$$\max_{j=1,\cdots,m-1} \frac{\sum_{i=1}^{j} V_i}{\sum_{i=1}^{j+1} V_i} < \gamma < 1$$

and let

$$\begin{aligned} \bar{P}(\lambda) &= (\lambda - \gamma)\bar{C}(\lambda) \\ &= -W_m\lambda^{2m} - [W_m + W_{m-1} - \gamma W_m]\lambda^{2m-1} \\ &\quad - \cdots - \left[\sum_{j=1}^{m} V_j - \gamma \sum_{j=1}^{m} W_j\right]\lambda^m - \left[\sum_{j=1}^{m-1} V_j - \gamma \sum_{j=1}^{m} V_j\right]\lambda^{m-1} - \cdots - \gamma V_m. \end{aligned}$$

By construction of γ, all coefficients of λ^p, $p \neq m$, are positive, and the coefficient of λ^m is negative. $\bar{P}(\gamma) = 0$. To see that there is another root $R > 1$, note that $\bar{P}(\lambda) > 0$ for λ real and sufficiently large. Furthermore, since $\bar{C}(1) = -C(1) < 0$ and $\gamma < 1$, $\bar{P}(1) < 0$. Hence $\bar{P}$ has another root $R > 1$. By Pellet's theorem, $\bar{P}(\lambda)$ has exactly m roots satisfying $|\lambda| \leqq \gamma$, and no more roots inside $|\lambda| = 1$. Thus $\bar{C}(\lambda)$ has exactly $m-1$ roots inside $|\lambda| = 1$.

If $\sum_{j=1}^{m} (jW_j + jV_j) < 0$, the proof is similar. □

Remark. Similar arguments are used in [11] to prove a refinement of this lemma. The conclusion also follows from the work of Berwald [13].

The previous two lemmas imply the following theorem.

THEOREM 3.1. *The linearization L_k of* (1.5) *around each $\Phi_k \in I$ defined by* (2.2) *is hyperbolic. Furthermore, the eigenvalues of the linearizations have absolute values that are bounded uniformly away from* 1.

Proof. The hyperbolicity is immediate from the previous lemmas. The uniformity holds provided that V_j, W_j, and $jW_j + jV_j$ are bounded away from zero. This is true by hypotheses H1 and H2 of § 2. □

We now wish to use Theorem 3.1 to show that there are solutions to (1.5) that stay arbitrarily close to any outer solution defined by (2.2) that satisfies $\Phi_k \in I$. This follows from a general theory involving exponential dichotomies for discrete systems [9]. For a linear difference equation

$$u_{k+1} = M_k u_k \tag{3.4}$$

let

$$\begin{aligned} t(k, l) &= M_{k-1} \cdots M_l \quad \text{if } k > l, \\ &= I \quad \text{if } k = l, \\ &= M_k^{-1} \cdots M_{l-1}^{-1} \quad \text{if } k < l \end{aligned}$$

be the transition matrix. Equation (3.4) has an exponential dichotomy on the natural numbers Z if there are positive constants K, α, and a family of projections P_k, $k \in Z$, such that

(i) $P_{k+1} M_k = M_k P_k$ for all $k \in Z$,

(ii) $|t(k, l) P_l| \leqq K e^{-\alpha(k-l)}$ for $l \leqq k$,

$|t(k, l)(I - P_l)| \leqq K e^{-\alpha(l-k)}$ for $l \geqq k$.

Thus, there is a family of projections that commutes with the operator of the equation. The projections are bounded and the solutions of (3.4) that lie in the range of $\{P_k\}$ decay exponentially, while those in the nullspace of $\{P_k\}$ grow exponentially. For $M_k \equiv \bar{M}$, $\bar{M}$ a hyperbolic matrix, it is easy to show that (3.4) has an exponential dichotomy on Z. Furthermore, perturbations of systems having exponential dichotomies also have exponential dichotomies [8], [9]. That is, if $M_k = \bar{M} + D_k$, and $|D_k|$ is sufficiently small uniformly in k, then (3.4) has an exponential dichotomy on Z. (See [9] for precise estimations on D_k and the constants for the exponential dichotomy for $\{M_k\}$ in terms of those of the constant coefficient case.)

Consider now the Taylor series expansion (3.2) of (1.5) around a solution $\{\Phi_k\}$ of (2.2). Written as a first-order system with

$$u_k = (x_{k,m-1}, \cdots, x_{k,1}, \eta_k, z_{k,1}, \cdots, z_{k,m-1}),$$

(1.5) has the form

$$u_{k+1} = L_k u_k + R(u_k) + r_k \tag{3.5}$$

where $R(u_k)$ is at least quadratic in components of u_k, and r_k is $0(1/N)$.

Equation (1.5) may be considered as a system for all $k \in Z$ by defining $\omega_k = \omega_1$ for $k \leqq 0$, $\omega_k = \omega_n$ for $k \geqq N+1$. If the frequencies $\{\omega_k\}$ are constant, then each outer solution $\{\Phi_k\}$ is also constant, and hence the linearization L_k of (1.5) around $\{\Phi_k\}$ is also constant. Thus, if the frequencies are sufficiently close to some constant, and the linearization associated with that constant is hyperbolic, the linear homogeneous system associated with (3.5) has an exponential dichotomy.

We wish to show that the outer solution $\{\Phi_k\}$, which in the above variables is $u_k \equiv 0$, is shadowed by a real solution of (3.5); more specifically, for N sufficiently large, there is a solution $\{\bar{u}_k\}$ of (3.5) such that $|\bar{u}_k|$ is $0(1/N)$ uniformly in k, $1 \leqq k \leqq N$. This will imply that there are a family of solutions $\{\bar{\phi}_k\}$ to (1.5), depending on N, such that $|\bar{\phi}_k - \Phi_k|$ is $0(1/N)$ as $N \to \infty$.

The result we need is a consequence of a general result for systems of the form (3.5) [9], which requires bounds on the $\{r_k\}$ and a Lipschitz constant for $R(u)$ in terms of the constants defining the exponential dichotomy for the associated linear system.

LEMMA [9]. *Let $\{M_k\}$ be a sequence of $n \times n$ matrices, $k \in Z$. Suppose that $u_{k+1} = M_k u_k$ has an exponential dichotomy on Z, with constants K, α, and projections P_k. Suppose that $G: R^n \to R^n$ satisfies $G(0) = 0$, and there are $\mu, \Delta > 0$, such that $|G_k(u) - G_k(\bar{u})| < \mu|u - \bar{u}|$ uniformly in k for $|u| < \Delta$, $|\bar{u}| < \Delta$. Let s_k be a sequence of n-vectors. Suppose further that*

$$\begin{aligned} &2K(1+e^{-\alpha})(1-e^{-\alpha})\mu \leqq 1, \\ &2K(1+e^{-\alpha})(1-e^{-\alpha})|s_k| \leqq \Delta. \end{aligned} \tag{3.6}$$

Then

$$u_{k+1} = M_k u_k + G(u_k) + s_k \tag{3.7}$$

has a unique solution $\{u_k\}$ such that $|u_k| < \Delta$ for all k. Moreover, $|u_k| \leqq 2K(1+e^{-\alpha})(1-e^{-\alpha}) \sup_{j \in Z} |s_j|$.

As discussed above, $\{L_k\}$, suitably extended, has an exponential dichotomy on Z, provided that $\{\omega_k\}$ are sufficiently close to some constant for which the associated L is hyperbolic. To apply the lemma to (3.5), we must show that $R(u)$ and $\{r_k\}$ satisfy the inequalities of (3.6). Indeed, we will show that they hold for N sufficiently large no matter what are the constants K and α associated with the exponential dichotomy of the linear homogeneous equation. We let $\Delta = 1/\sqrt{N}$. Since each r_k is $0(1/N)$, uniformly in k (even in the extended system), the second inequality of (3.6) holds for N sufficiently large. Since $R(u)$ is at least quadratic in the components of u, there is a Lipschitz constant μ for R that is $0(1/\sqrt{N})$ if $u, \bar{u}$ are restricted to have norm less than or equal to Δ. It follows that the first inequality of (3.6) also holds for N sufficiently large.

We have now established the following theorem.

THEOREM 3.2. *Suppose $\Omega, \{\omega_K\}, H^{\pm}$ satisfy* H1 *and* H2, *and $\{\omega_k\}$ are sufficiently close to a constant so that $\{L_k\}$ of* (3.5) *has an exponential dichotomy. Then there is a solution to* (1.5) *that shadows the solution to* (2.2) *within* $0(1/N)$. □

4. On the structure of the solutions for a finite chain. As seen in §§ 2 and 3, for each Ω in some range, (1.5) may be thought of as a first-order system of equations of dimension $2m-1$. By extending this system as in § 3, we may think of (1.5) as defined for all k. For each Ω in some region, there is a solution to (2.2) that is constant for $k<0$ and $k>N$, and a unique solution $\{\phi_k^\Omega\}$ to (1.5) that shadows it for all k.

We have seen that, with suitable restrictions on $\{\omega_k\}$ and Ω, the linearization of (1.5) around a solution $\{\Phi_k^\Omega\}$ to (2.2) is uniformly hyperbolic, and the linear difference equation (3.1) has an exponential dichotomy. Since $\phi_k^\Omega - \Phi_k^\Omega$ approaches zero uniformly in k as $N\to\infty$, by the perturbation property ("roughness") of exponential dichotomies [8], [9], the linearization of (1.5) around $\{\phi_k^\Omega\}$ also possesses an exponential dichotomy. (Note that the equations analogous to (3.5), but linearized around $\{\phi_k^\Omega\}$ instead of $\{\Phi_k^\Omega\}$, have no inhomogeneous term r_k.) It is then possible to construct a stable manifold for each such $\{\phi_k^\Omega\}$; although stable manifolds are usually defined for fixed points of an autonomous dynamical system, the usual definition and usual proof are essentially unchanged if the fixed point is replaced by a compact orbit of a nonautonomous system and the linearization around that orbit possesses an exponential dichotomy on Z. The dimension of this stable manifold is the number of eigenvalues of the linearization having absolute value less than 1. For definiteness, we assume that we have an outer solution $\{\phi_k^\Omega\}$ for which this number is $m-1$ rather than m. (Similar arguments will work for the other case.)

A finite chain differs from an infinite chain described by extensions of (1.5) by having modified equations at each of the ends. The most straightforward way to modify the equations is to remove all terms that involve nonexistent oscillators. That is, if there are N oscillators whose phases are denoted by $\theta_1,\cdots,\theta_N$, then only terms involving $\phi_1,\cdots,\phi_{N-1}$ are retained in the equations. For $m=1$, this leads to (1.6). For a chain coupled to m neighbors on each side, it involves a modification of the first m and the last m equations. A modification of the $k=j$ equation of (1.5), $1\leqq j\leqq m$, is a restriction on $\phi_1,\cdots,\phi_{m+j-1}$. We may think of these m restrictions as confining the variables $\eta_m, x_{m,i}, z_{m,i}$ to an $m-1$-dimensional submanifold M_I of its $2m-1$-dimensional space. Similarly, modifications of the last m equations define an $m-1$-dimensional submanifold M_E of $\eta_{N-m}, x_{N-m,i}, z_{N-m,i}$.

The structure of the resulting boundary value problem is very close to that of the classical Sil'nikov problem [10], [14]. In the latter, an ordinary differential equation with a hyperbolic critical point p is given. At $t=0$, the solution is required to lie on a given manifold transverse to the stable manifold of the critical point and of complementary dimension; at $t=t_1$, the solution is required to be on another given manifold transverse to the unstable manifold of the critical point and of complementary dimension. Then if t_1-t_0 is sufficiently large (the bound depending on the constants associated with the exponential dichotomy of the linearized system around the critical point, and the size of the given manifolds), there is a unique solution with the required properties (see Fig. 4.1). As $t_1-t_0\to\infty$, the solution approaches the constant solution p except for boundary layers at each end.

The result required here is different from the Sil'nikov problem in that we deal with discrete nonautonomous systems instead of continuous autonomous ones. Furthermore, there is a one-parameter family of systems instead of a single system. However, as long as there are exponential dichotomies for the nonautonomous systems, the idea is essentially the same. The dimensions are required to be such that the following hypothesis holds.

$\bar{\text{H}}1$. Either the stable manifold $W^s(\Omega)$ intersects the manifold $M_I(\Omega)$ of initial conditions for only one value of the parameter or the unstable manifold intersects the

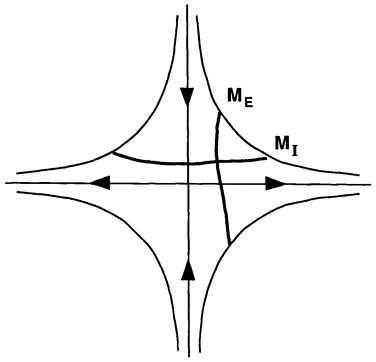


FIG. 4.1. *The geometric solution to the Sil'nikov problem. The manifold M_I of conditions at $t = t_0$ is swept forward under the flow and becomes C^1 close to the unstable manifold of p. The manifold M_E of conditions at $t = t_1$ is swept backwards under the flow and becomes C^1 close to the stable manifold. For $t = (t_1 - t_0)/2$, the images intersect in a point Q whose trajectory is the required solution.*

manifold $M_E(\Omega)$ of end conditions at exactly one value of the parameter (see Fig. 4.2a-c).

Let Ω_0 denote the above value of the parameter and, for the sake of definiteness, assume that it is determined as the intersection of M_I with the stable manifold corresponding to Ω_0. We require the next hypothesis.

$\bar{\text{H}}2$. For each Ω in a neighborhood of Ω_0, the unstable manifold $W^u(\Omega)$ has a nonzero transverse intersection with $M_E(\Omega)$ (see Fig. 4.3).

Under hypotheses $\bar{\text{H}}1$ and $\bar{\text{H}}2$ there is a locally unique solution to the boundary value problem. Theorem 4.1 below states this in a more general form that allows the parameter space to be of any dimension. A similar result is true if Ω_0 is determined at the opposite end of the interval.

THEOREM 4.1. *Let*

$$u_{k+1} = A_{k,\Omega}u_k + R_{k,\Omega}(u_k) \equiv T_\Omega(u_k), \tag{4.1}$$

$u \in R^n$, be a parameterized family of difference equations, where A and R depend smoothly on the parameter $\Omega \in R^j$ and $R_{k,\Omega}$ is at least quadratic in u and uniformly bounded in k. Suppose that for each value of the parameter in some range, (4.1) *has an exponential dichotomy on Z, with the families of projections varying smoothly with Ω. Let p and q be the dimensions of the contracting and expanding subspaces, respectively, so $p + q = n$. By the "center stable manifold" of dimension $p + j$ we mean the union of the stable manifolds of* (4.1) *over all relevant Ω_i, and similarly for the unstable manifold. Suppose further that at $k = k_1$, the solution is required to lie on an $n - p - j$-dimensional manifold $M_I(\Omega)$ and that the intersection in $R^n \times R^j$ of the center stable manifold with $M_I = \cup_\Omega M_I(\Omega)$ is a unique point Q. Also, at $k = k_N$, the solution is required to lie on an $n - q$-dimensional manifold $M_E(\Omega)$ satisfying $\bar{\text{H}}2$. Then for $k_N - k_1$ sufficiently large, there is a unique solution $\{u_k\}$ satisfying $u_k \to 0$ as $k_1 - k_N \to \infty$, except possibly in regions of size $0(1/\sqrt{N})$ around $k = k_1$ and $k = k_N$. In that limit, the solution approaches Q at $k = k_1$ and, except for a boundary layer near $k = k_N$, the solution approaches the center stable manifold.*

Proof. We first choose a set of coordinates in which the calculations become easier. Fix Ω. In the presence of an exponential dichotomy, a k-dependent change of coordinates may be made so that the new linear part $\bar{A}_{k,\Omega}$ is block diagonal, with the upper left-hand block of size $p \times p$ having eigenvalues with absolute value less than 1, and the lower right-hand block is of size $q \times q$ having eigenvalues with absolute value

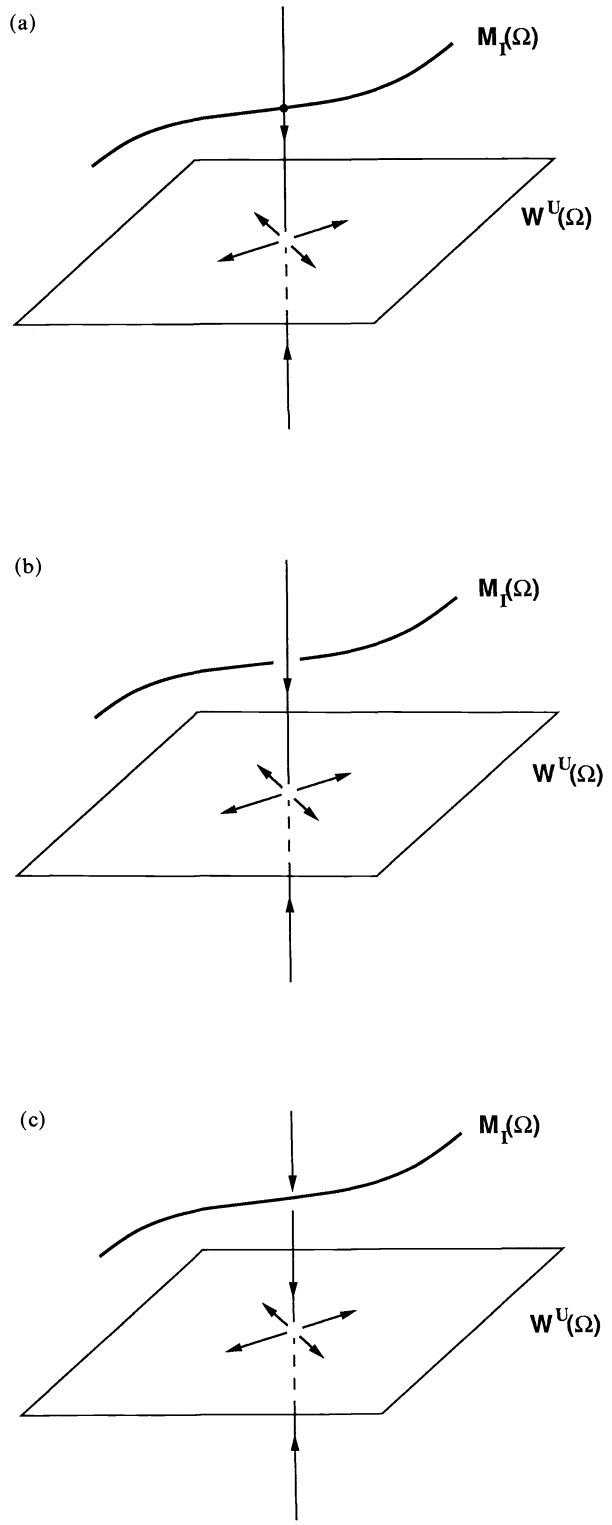


FIG. 4.2. *The dynamics at* $k = k_1$ *and the manifold* $M_I(\Omega)$ *for* (a) $\Omega = \Omega_0$, (b), (c) $\Omega \neq \Omega_0$. *In this figure,* $j = 1$ *and* dim $M_I(\Omega)$ *is one less than the codimension of* $W^s(\Omega)$.

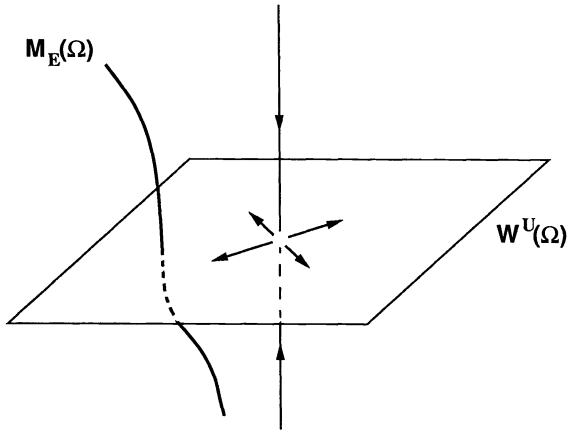


FIG. 4.3. *The dynamics at $k = k_N$ for Ω near Ω_0. $M_E(\Omega)$ has a transverse intersection with $W^u(\Omega)$.*

greater than 1 [8]; this can be done in a way that is smooth with respect to Ω. (The proof in [8] is for continuous equations, but works as well for discrete systems.) Without loss of generality, we may now assume that $A_{k,\Omega}$ is of this form.

For each Ω, the $n-p$ dimensional manifold M_I is transverse to the stable manifold $W^s(\Omega)$. It follows from the λ-lemma [16] that for k sufficiently large, the image under T_Ω^k of M_I is arbitrarily C^1 close to $W^u(\Omega)$, along with its first derivatives. Similarly, the $n-q$-dimensional manifold $M_E(\Omega)$ is transverse to $W^u(\Omega)$, so by the λ-lemma, its image under T_Ω^{-k} for sufficiently large k is arbitrarily C^1 close to $W^s(\Omega)$. (Note: the λ-lemma is usually stated for autonomous systems. However, the proof remains valid for nonautonomous systems (4.1) provided that the contracting and expanding subspaces are independent of k and the eigenvalues of $A_{k,\Omega}$ are bounded away from the unit circle uniformly in k.) Thus, for N sufficiently large, the image of M_I at $k=[N/2]$ transversely crosses the image of $M_E(\Omega)$ at $k=[N/2]$. (Here $[\cdot]$ denotes "least integer in.") Since this is true for each Ω, the intersections determine a j-dimensional space in $R^n \times R^j$, parameterized by Ω.

On this j-dimensional manifold, each point is assigned a parameter Ω in an additional way. In other words, by construction, it is the intersection of $M_I = \bigcup_\Omega M_I(\Omega)$ with the manifold $M_E(\Omega)$, and hence occurs on $M_I(\Omega)$ for some Ω. Thus there are two maps from a region of R^j to R^j given by these two assignments. Denote the first by α and the second by β. The required solution must satisfy $\alpha(\Omega)=\beta(\Omega)$.

By construction, $\alpha(\Omega)=\Omega$. To estimate β, we note that by the λ-lemma, the backward image of any manifold transverse to $W^u(\Omega)$ (in particular $M_E(\Omega)$ for any Ω) approaches $W^s(\Omega)$, and hence intersects M_I at a point approximately independent of Ω. This implies that $d\beta/d\Omega$ has a small Jacobian, and hence, by the Implicit Function Theorem, there is a unique Ω, depending on N, such that $\alpha(\Omega)=\beta(\Omega)$. This solution has the limiting properties (as $N\to\infty$) stated in the theorem. □

Remark. If $m=1$, under some restrictions on $\{\omega_k\}$ there is exactly one solution to the BVP (2.5), (1.6). In particular, either the outer solution satisfying $H^-(-\phi_1)=0$ can be matched to a boundary layer near $k=N$ *or* the outer solution satisfying $H^+(\phi_N)=0$ can be matched to a boundary layer near $k=1$. In [3], there are explicit and easily checked formulas to determine which side of the chain has the boundary layer, and what is the determining frequency Ω.

For $m>1$, we do not have such formulas, nor do we yet have a uniqueness result. Furthermore, there are generally boundary layers at both ends. However, there is still

some structure that helps us to compute our solution without fully solving the equation numerically. Let $\mathbf{O}^+$ (respectively, $\mathbf{O}^-$) denote the set of outer solutions satisfying H1 and H2 along which $F'>0$ (respectively, $F'<0$). By Lemma 3.2, on $\mathbf{O}^+$ (respectively, $\mathbf{O}^-$) the stable manifold has dimension m (respectively, $m-1$) and the unstable manifold has dimension $m-1$ (respectively, m). Thus, the center stable manifold at $k=m$ corresponding to $\mathbf{O}^-$ intersects an open set of initial manifolds M_I at a single point. Similarly, the center unstable manifold at $k=N-m$ corresponding to $\mathbf{O}^+$ intersects an open set of manifolds M_E at a point. Suppose that the other boundary manifold (M_E if $F'<0$, M_I if $F'>0$) intersects the outer solution corresponding to that point. By Theorem 4.1, for N large there is a solution to (1.5) that passes close to that point. If the value of Ω at that point can be computed, an approximate solution is easily found from (2.2).

Suppose for definiteness we consider $\mathbf{O}^-$. Due to the nonlinearity of the boundary layer, it is not easy, given any set of initial conditions, to compute the intersection of the initial conditions and the center stable manifold. Indeed, our abstract proof is for an open set of initial conditions and does not necessarily apply to any particular set. However, the exponential decay property of trajectories on a stable manifold allows us to compute a plausible sequence of approximations to that intersection. As an example, consider (1.5) modified as described above to omit all terms containing θ_k for $k\notin[1, N]$. Also, for simplicity, assume that ω_k is independent of k. The approximation procedure is as follows: For any $j\geqq 1$, replace all ϕ_k, $k>j$, in the modified (1.5) by an unknown $\bar{\phi}$ independent of k. (The approximation assumes that the trajectory "reaches" the outer solution, which is constant, in $j+1$ steps.) Then the first $j+2$ equations of the modified (1.5) have in them $j+2$ unknowns $\phi_1, \cdots, \phi_j, \bar{\phi}, \Omega$. These

TABLE 1
$m=2$.

j	$\bar{\phi}$	Ω
2	−.46365	−.59544
3	−.23103	−.41889
4	−.31779	−.46224
5	−.27365	−.45055
6	−.29339	−.45382
7	−.28468	−.45297
8	−.28848	−.45320
true	−.28733	−.45315

TABLE 2
$m=5$.

j	$\bar{\phi}$	Ω
2	−.46365	−.71250
4	−.15411	−.90279
5	−.11566	−.82677
6	−.09261	−.78427
8	−.14142	−.85154
10	−.12797	−.84314
20	−.13314	−.84534
true	−.13314	−.84534

$j+2$ equations may be solved independently of all of the other equations. If $m=1$, it may be checked that this procedure with $j=1$ produces $\phi_1=\bar{\phi}$ and $\Omega=\omega_1+2f(\phi_R)$, which is the correct answer for nearest-neighbor coupling. We have not been able to show that roots exist for this approximation, nor have we been able to show that it converges to the correct roots. Nevertheless, numerical solutions to these appear to converge to a unique root that is in good agreement with the frequency obtained by integrating the full equations. For $m=2$, we need only solve the reduced system up to about $j=8$, and in Table 1 we show the results of a typical example. The advantages of the approximation are most evident for m small; indeed, we need solve only eight equations instead of $N \gg m$ differential equations. For m larger, the approximation takes longer to converge; $m=5$ requires $j \approx 10$–20.

5. Nearest-neighbor coupling vs. multiple coupling. Away from the boundaries, the solution to (1.5) with modified end conditions behaves like one of a family of outer solutions, at least for an open set of boundary conditions. Thus, we first compare outer solutions (2.2) with those of the case $m=1$ given by (2.3).

One effect of coupling m neighbors is to strengthen the coupling and thereby diminish the effects of any possible frequency differences. That is, for fixed Ω, the solution $\{\Phi_k\}$ to (2.2) or (2.3) deviates from being constant if $\{\omega_k\}$ is not constant. In (2.3), the deviation of the $\{\Phi_k\}$ from constancy depends as well on the size of f'; doubling f leads to the same solution as halving Ω and each ω_k (and thus halving the total frequency difference). In (2.2), the dependence is similar. That is, $F'(\phi)=f_1'(\phi)+2f_2'(2\phi)+\cdots+mf_m'(m\phi)$; provided the solution stays in the region in which $f_j'(j\phi)$ all have the same sign, $|F'(\phi)|$ grows with increasing m, and thus a fixed total frequency difference in the $\{\omega_k\}$ produces a smaller difference in the $\{\Phi_k\}$.

The multiple coupling has another important consequence besides lessening the effects of frequency differences, a consequence that shows up even when ω_k is independent of k, and even if the coupling strength is normalized in such a way that the F obtained from (2.2) has the same magnitude as the f in (2.3) (e.g., the difference between the maximum and minimum values of F is the same as that of f). More specifically, *the phase differences* $\{\Phi_k\}$ *decrease as* M *is increased.* This is most easily seen by taking a simple isotropic example $H_j^{\pm}(\theta)=c(m,\rho)\rho^{j-1}(\sin(\theta)+\alpha\cos(\theta))$, where $c(m,\rho)=\gamma(\rho-1)/(\rho^m-1)$, $\rho \leqq 1$, $\gamma>0$ is the normalization. This allows us to let the strength of coupling decrease as the distance between oscillators increases. If $\rho=1$, then $c(m,\rho)=1/m$. We may explicitly compute $F(\phi)$ by summing the series; we obtain

$$F(\phi)=\alpha c(m,\rho)\frac{\rho^{m+1}\cos(m\phi)-\rho-\rho^m\cos((m+1)\phi)+\cos(\phi)}{\rho^2+1-2\rho\cos(\phi)}. \tag{5.1}$$

A similar but more complicated expression arises in the anisotropic case, e.g., one in which the strength is different for H_j^+ and H_j^-. In Fig. 5.1, we sketch $F(\phi)$ for various values of m. As can easily be seen, the function $F(\phi)$ is much *narrower* than $f(\phi)$ although it has the same amplitude. Furthermore, if $\phi \in (-\pi/2m, \pi/2m)$, $f_j'(j\phi)$ and $F'(\phi)$ have the same sign. With normalized coupling strengths as above, both forms of coupling (multiple and nearest neighbor) support similarly-sized frequency differences, but the resultant phase differences are considerably smaller for the multiply coupled case. For comparison, in Fig. 5.2 we sketch an example of $F(\phi)$ with anisotropy ($\alpha=-.5$, $\gamma^+=.6$, $\gamma^-=1.0$) that shows the same narrowing effect.

Note that as the number of neighbors increases the function F becomes more narrow, even if normalized to keep the same amplitude. In particular, as $m \to \infty$, and

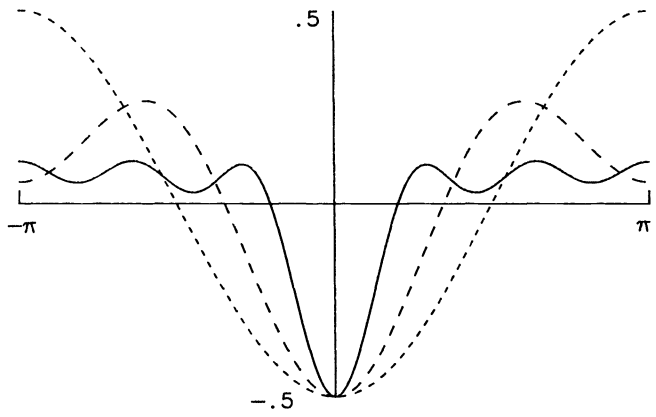


FIG. 5.1. *$F(\phi)$ as a function of ϕ for three values of m and $f(\phi) = -.5 \cos \phi$ (isotropic medium). The short dashes trace the $m = 1$ curve, the long dashes $m = 2$, and the solid line $m = 3$.*

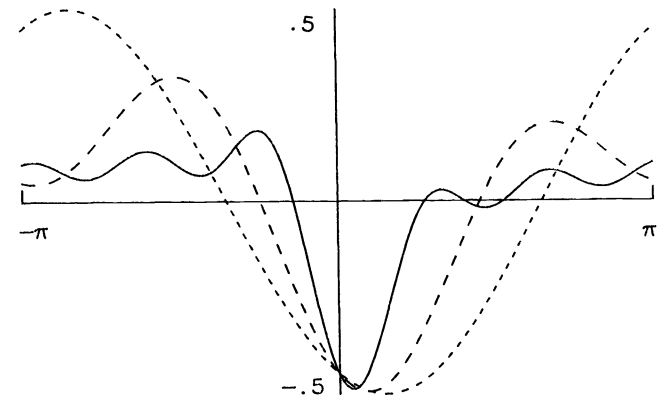


FIG. 5.2. *$F(\phi)$ as a function of ϕ for $f(\phi) = -.4 \cos \phi - .2 \sin \phi$ (anisotropic medium), and $m = 1, 2, 3$. The dashed and solid curves have the same meaning as above.*

$\rho = 1$, $F(\phi)$ in (5.1) tends to a function that is zero everywhere except at the origin, where it is 1. Thus, for large m, the function F tends to a narrower and narrower peaked function. (We still consider m small relative to the number of oscillators.) For a fixed number of neighbors, $F(\phi)$ becomes wider as ρ decreases. As $\rho \to 0$, m fixed, and with the amplitudes normalized, $F(\phi) \to f(\phi)$, the nearest neighbor case. For $\rho < 1$, as $m \to \infty$, $F(\phi)$ tends to a peaked function that is narrower in width than $f(\phi) = \cos(\phi)$.

It follows from the narrowness (i.e., the larger size of F' compared with f') that changes in ω_k produce smaller changes in phase differences for F than for f. We now give simulations to show how the effects of frequency differences are diminished by multiple coupling even when F is normalized. In the next several figures, we fix the frequency gradient and compare the phase differences for one-neighbor, two-neighbor, and five-neighbor coupling. We normalize so that the amplitudes of $F(\phi)$ are equivalent. In Fig. 5.3, we show the results of isotropic coupling with $\rho = 1$ (no damping in the coupling strengths) for $m = 1, 2$, and 5. ($H_j^{\pm}(\theta) = \sin(\theta) - .5 \cos(\theta)$.) The chain has a linear frequency gradient from $\omega_1 = 1$ to $\omega_{100} = 0.5$. There is very little difference in the general *shapes* of the phase-difference curves, but the *magnitude* of the curves are quite different, as are their slopes; clearly, the slope is smallest for $m = 5$, so the effect

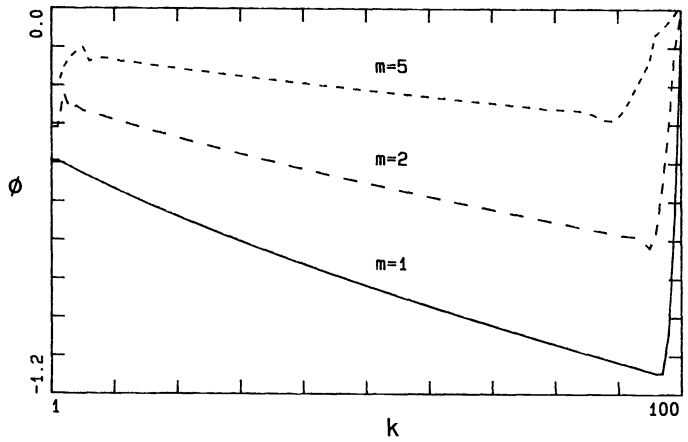


FIG. 5.3. ϕ_k *as a function of* k *in an isotropic medium with a frequency gradient*: $\omega_k = 1 - .005k$, $\rho = 1$, $k = 1, \cdots, 100$. *The graph is drawn for* $m = 1, 2, 3$ *as in Figs.* 5.1 *and* 5.2.

of the frequency gradient is smaller. Note that even with a large frequency gradient, the phase differences at each point away from the boundary stay between zero and $-\pi/2m$, i.e., in the region for which $f_j'(j\phi)$ all have the same sign as $F'(\phi)$. The phase-locked frequency of the chain is similar in the three cases (0.098, 0.087, and 0.066 for $m = 1, 2$, and 5, respectively). One curious effect is that the boundary layer on the right is sharpest for the nearest-neighbor case and gets coarser as m increases.

Figure 5.4 depicts the phase differences for different values of ρ for five neighbors. All parameters except ρ are as in the previous figure. As we derived from (5.1) above, the magnitude of the phase differences is larger for $\rho = 0.8$ and smaller for $\rho = 1.25$. Small amounts of anisotropy make little difference in the qualitative picture when there is an imposed frequency gradient; in the absence of such a gradient, differences from the isotropic case are dramatic (see, e.g., [2]).

Although the mathematical results in this paper assume that the frequencies in the chain slowly vary, a similar argument could be formally applied to the situation in which the frequencies are close but randomly distributed. Nearest-neighbor coupling

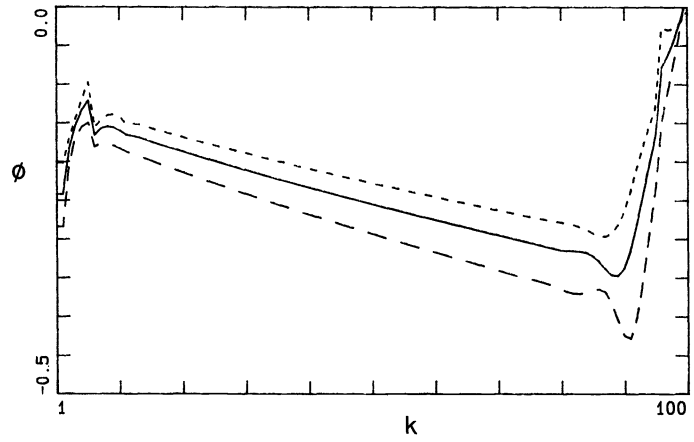


FIG. 5.4. ϕ_k *as a function of* k *in an isotropic medium,* $m = 5$, *three values of* ρ *and other parameters as in Fig.* 5.3. *Long dashes trace the* $\rho = .8$ *curve, short dashes* $\rho = 1.25$, *and the solid curve* $\rho = 1$. (*Note that* $\rho = 1.25$ *means that distant neighbors have stronger interactions than nearest neighbors.*)

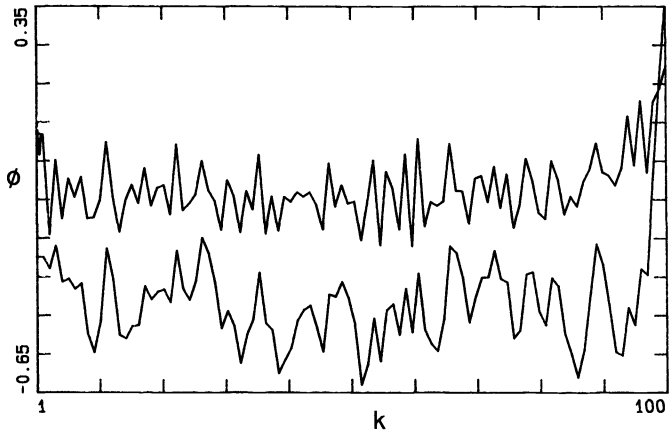


FIG. 5.5. *ϕ_k versus k for an anisotropic medium, with ω_k chosen randomly from the interval* [.7, 1]. *The lower graph is $m = 1$ and the upper is $m = 5$. $H^+(\phi) = .8H^-(\phi)$, $H^-(\phi) = -.5\cos\phi + \sin\phi$.*

does not buffer as well against these effects as the multiply coupled situation. This result is shown in Fig. 5.5, where we depict the phase differences for a chain whose frequencies are randomly chosen from the interval (.7, 1). We compare coupling to nearest neighbors with coupling to five nearest neighbors. ($H^+(\phi) = .8(\sin(\phi) - .5\cos(\phi))/m$, $H^-(\phi) = (\sin(\phi) - .5\cos(\phi))/m$, $m = 1$ or 5.) The multiple-coupling case shows an average phase shift that is closer to zero than the nearest neighbor case ($\langle\phi\rangle = -.119$ for five neighbors versus $\langle\phi\rangle = -.409$ for nearest neighbors). The variation is also less for the multiply coupled case ($\sigma_5 = .087$ vs. $\sigma_1 = .136$). If we repeat the calculation for mean and standard deviations ignoring the last five oscillators (where there is a boundary layer), similar quantitative results are found ($\langle\phi\rangle_5 = -.130$, $\sigma_5 = .071$; $\langle\phi\rangle_1 = -.428$, $\sigma_1 = .090$).

Acknowledgment. We thank S. Laederich for helpful conversations about exponential dichotomies. Some of this work was done as part of the second author's Ph.D. thesis at Northeastern University, in 1988.

REFERENCES

[1] G. B. ERMENTROUT AND N. KOPELL, *Frequency plateaus in a chain of weakly coupled oscillators*, I, SIAM J. Math. Anal., 15 (1984), pp. 215-237.

[2] N. KOPELL AND G. B. ERMENTROUT, *Symmetry and phaselocking in chains of weakly coupled oscillators*, Comm. Pure Appl. Math., 39 (1986), pp. 623-660.

[3] ———, *Phase transitions and other phenomena in chains of coupled oscillators*, SIAM J. Appl. Math., 50 (1990), to appear.

[4] N. KOPELL, *Coupled oscillators and the design of central pattern generators*, Math. Biosci., 90 (1988), pp. 87-109.

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

[6] G. B. ERMENTROUT AND N. KOPELL, *Oscillator death in systems of coupled neural oscillators*, SIAM J. Appl. Math., 50 (1990), pp. 125-146.

[7] ———, *Multiple pulse interactions and averaging in systems of coupled neural oscillators*, J. Math. Biol., to appear.

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

[9] K. J. PALMER, *Exponential dichotomies, the shadowing lemma and transversal homoclinic points*, Dynamics Reported, 1 (1988), pp. 265-306.

[10] BO DENG, *Sil'nikov problem, exponential expansion, strong λ-lemma, C^1 linearization and homoclinic bifurcation*, Lefschetz Center for Dynamical Systems, Brown University, Providence, RI, 1988, preprint.

[11] M. MARDEN, *The Geometry of the Zeroes*, Math. Surveys III, American Mathematical Society, Providence, RI, 1949.

[12] M. A. PELLET, *Sur un mode de separation des racines des equations et la formule de Lagrange*, Bull. Sci. Math., 5 (1881), pp. 393–395.

[13] L. BERWALD, *Über einige mit dem Satz von Kakeya verwandte Satze*, Math. Z., 37 (1933), pp. 61–76.

[14] L. P. SIL'NIKOV, *The existence of a denumerable set of periodic motions on the four-dimensional space in an extended neighborhood of a saddle focus*, Soviet Math. Dokl., 8 (1967), pp. 54–57.

[15] S. LAEDERICH, *Boundary value problems for partial differential equations with exponential dichotomies*, J. Differential Equations, to appear.

[16] J. PALIS AND W. DEMELO, *Geometric Theory of Dynamical Systems*, Springer-Verlag, Berlin, New York, 1982.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 954-979, July 1990
© 1990 Society for Industrial and Applied Mathematics
008

SUBHARMONIC BRANCHING IN REVERSIBLE SYSTEMS*

A. VANDERBAUWHEDE†

Abstract. The branching of subharmonic solutions at a symmetric periodic solution of an autonomous reversible system is studied. Here "symmetric" means "invariant under time reversal." It is shown that generically each such symmetric periodic solution belongs to a one-parameter family of similar periodic solutions. Along such a family solutions having multipliers that are roots of unity can be met generically. It is shown that at such solutions further branching of subharmonic solutions will generically occur.

Key words. periodic solutions, reversible systems, period-doubling, subharmonic branching, Lyapunov-Schmidt method

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

1. Introduction and preliminaries. In this paper we will consider autonomous systems of the form

$$\dot{x} = f(x) \tag{1.1}$$

satisfying the following hypotheses:

(H0) (i) $x \in \mathbf{R}^{2n}$ and $f: \mathbf{R}^{2n} \to \mathbf{R}^{2n}$ is sufficiently smooth.

(ii) There exists some $R \in \mathscr{L}(\mathbf{R}^{2n})$, with $R^2 = I$ and $\dim \operatorname{Fix}(R) = n$, such that

$$f(Rx) = -Rf(x) \quad \forall x \in \mathbf{R}^{2n}. \tag{1.2}$$

Some of the results that we will state further on require only a finite smoothness for f (i.e., $f \in C^k$ for some sufficiently large $k < \infty$), whereas others, in particular those on cascades of subharmonics, require f to be C^∞-smooth.

In (H0)(ii) we have used the following notation: If X is any Banach space and $\mathscr{A} \subset \mathscr{L}(X)$, then we define

$$\operatorname{Fix}(\mathscr{A}) := \{x \in X \mid Ax = x, \ \forall A \in \mathscr{A}\}. \tag{1.3}$$

The hypothesis (H0)(ii) is satisfied in a natural way for *oscillation systems*; these are systems of the form

$$\ddot{y} = g(y), \tag{1.4}$$

with $y \in \mathbf{R}^n$ and $g: \mathbf{R}^n \to \mathbf{R}^n$. System (1.4) reduces to (1.1) by putting

$$x = \begin{pmatrix} y \\ \dot{y} \end{pmatrix}, \qquad f(x) = \begin{pmatrix} x_2 \\ g(x_1) \end{pmatrix} \quad \text{for } x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

$$R = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}.$$

Condition (1.2) implies that if $x(t)$ is any solution of (1.1), then so is $\tilde{x}(t) := Rx(-t)$. For this reason we call systems satisfying (H0) *reversible* (or more precisely, time-reversible) systems.

* Received by the editors March 11, 1989; accepted for publication (in revised form) August 2, 1989. Part of this work was done while the author was visiting at the University of Nice, Nice, France.

† Instituut voor Theoretische Mechanika, Rijksuniversiteit Gent, Krijgslaan 281, B-9000 Gent, Belgium.

The condition $R^2 = I$ implies that $\{I, R\}$ forms a (compact) group of linear operators acting on $\mathbf{R}^{2n}$. By a general result (see, e.g., Bredon [1] or Vanderbauwhede [9]) we may then assume that this action is orthogonal, i.e., $RR^T = I$, and hence also $R = R^T$. The starting point of our analysis will be a nonconstant symmetric periodic solution of (1.1); more precisely, we assume the following hypothesis:

(H1) (i) Equation (1.1) has a nonconstant periodic solution $x_0(t)$, with minimal period $T_0 > 0$, satisfying

$$Rx_0(-t) = x_0(t+\phi) \quad \forall t \in \mathbf{R} \tag{1.5}$$

for some $\phi \in \mathbf{R}$.

Later we will complement hypothesis (H1) with a second part, (H1)(ii). Condition (1.5) means that the companion solution $Rx_0(-t)$ of $x_0(t)$ has the same orbit as $x_0(t)$ itself; we call such solutions *symmetric*. By shifting the origin of time (i.e., by replacing t by $t - \phi/2$ in this case) we may, without loss of generality, assume that $\phi = 0$ in (1.5), and hence

$$Rx_0(-t) = x_0(t) \quad \forall t \in \mathbf{R}. \tag{1.6}$$

The problem we want to discuss in this paper is then the following: Given an integer $q \geqq 1$, describe all periodic orbits of (1.1) with the properties that:

(a) Their orbits lie in a neighborhood of the orbit of $x_0(t)$, which we denote by

$$\kappa_0 := \{x_0(t) \mid t \in \mathbf{R}\}. \tag{1.7}$$

(b) They have a period near qT_0.

The case $q = 2$ corresponds to *period-doubling*, whereas for $q \geqq 3$ we will talk about *subharmonic branching*.

In § 2 we use a geometrical argument to obtain a result due to Devaney [2], namely, that generically $x_0(t)$ will belong to a one-parameter family $\{x_\alpha(t) \mid \alpha \in \mathbf{R}\}$ of symmetric and periodic solutions of (1.1); a consequence of this result will motivate part (ii) of our hypothesis (H1). From § 3 on we study our problem in the form of an abstract equation suitable for a treatment by a Lyapunov-Schmidt method. In § 3 we study the linearization of this equation. In § 4 we make a reduction of the abstract problem; this reduction forms the counterpart (in our abstract setting) of the classical Poincaré map in phase space; as a result we find a classical bifurcation problem with D_q-symmetry. In §§ 5–7 we study this problem for the cases $q = 1$, $q = 2$, and $q \geqq 3$, respectively. In the case $q = 1$, we recover the solution branch $\{x_\alpha\}$ mentioned before. For the cases $q = 2$ and $q \geqq 3$, we obtain conditions that ensure the existence near x_0 of further branches of symmetric periodic solutions with period near qT_0. For $q = 2$ we find a single branch of symmetric periodic orbits with double period, branching at x_0 from the primary branch $\{x_\alpha\}$; for $q \geqq 3$, two such branches of subharmonics bifurcate at x_0 from the primary branch. Our approach is a combination of a reduction technique introduced by Vanderbauwhede in some earlier papers [6], [7], with the bifurcation analysis made in [8] for the problem of subharmonic bifurcation in reversible but nonautonomous equations. For the case $q \geqq 5$, results similar to the ones described here have been obtained by Sevryuk [5], using the Poincaré map and normal form theory.

The results proved in this paper show that reversible systems can have a rich variety of periodic solutions. Generically, the mere existence of a symmetric periodic solution implies the existence of a whole branch of such solutions, and if along such branch some of the multipliers get trapped on the unit circle, then branching of

subharmonic solutions will occur. In our theorems we fix some q to isolate particular solutions, but of course the complete picture is much more involved. Consider, for example, the case of period-doubling (Theorem 3): Our transversality condition (H2)(ii) implies that along the primary branch $\{x_\alpha\}$ we will have for one sign of α simple Floquet multipliers on the unit circle, travelling with nonzero speed. These multipliers will pass through an infinity of roots of unity (with high values of q), and at each of these give rise to subharmonic branching, according to Theorem 4 or Theorem 5. Probably the neighborhoods in which our local analysis is valid will shrink down to a point as $\alpha \to 0$ and $q \to \infty$. Moreover, the solutions along the double-period branch will have a pair of multipliers near $+1$; we have not yet tried to calculate these multipliers, but it is conceivable that under certain conditions these multipliers will be on the unit circle, giving rise to further subharmonic branching. A similar complexity arises in the case $q \geqq 3$. The transversality condition (Hq)(iii) implies that the multipliers along the primary branch will meet an infinity of roots of unity, at each of them giving rise to an appropriate subharmonic branching. Moreover, the calculations that we made for nonautonomous reversible systems (see [8]) suggest that along part of the subharmonic branch there will be a pair of simple multipliers on the unit circle (near one), giving rise to further subharmonic branching. If this is indeed the case, then we can expect whole cascades of subharmonic branchings, with ever increasing periods. It seems to be an interesting problem to see whether the approach of this paper can be extended to capture such phenomena in a more global way.

Our results should be compared with what happens in generic nonreversible systems. In those, periodic solutions are isolated, and we can see period-doubling only by changing a parameter; for subharmonic bifurcation we even need two parameters. (Remember that our system (1.1) does not depend on any explicit parameter.) There is, however, another particular class of systems whose set of periodic solutions seems to behave in very much the same way as what we find here for reversible systems: the results obtained by Meyer [4] for generic Hamiltonian systems show a similar structure of branching families of subharmonic solutions. We should also remark that although the two classes do not coincide still many Hamiltonian systems are also time reversible. For a general introduction to reversible systems and their relation to Hamiltonian systems, refer to the paper of Devaney [2] and to the recent lecture notes of Sevryuk [5].

2. A continuation result. In this section we use a Poincaré map to prove the following continuation result (see Devaney [2]).

THEOREM 1. *Generically the orbit κ_0 of $x_0(t)$ belongs to a one-parameter family of periodic orbits of* (1.1), *all corresponding to symmetric periodic solutions, and with the minimal period changing smoothly along the family.*

Proof. The proof is based on the following simple observation. Suppose that $x(t)$ is a solution of (1.1) such that

$$Rx(0) = x(0) \quad \text{and} \quad Rx\left(\frac{T}{2}\right) = x\left(\frac{T}{2}\right) \tag{2.1}$$

for some $T > 0$. Then the first condition of (2.1) implies that $Rx(-t) = x(t)$ for all $t \in \mathbf{R}$ (i.e., x is a symmetric solution), whereas from the second condition we find

$$x\left(-\frac{T}{2}\right) = Rx\left(\frac{T}{2}\right) = x\left(\frac{T}{2}\right).$$

We conclude that (2.1) implies that $x(t)$ is symmetric and T-periodic.

Now let $x_0(t)$ be the solution given by (H1)(i) and satisfying (1.6). At the point $x_0(0)$ we construct a transversal section Σ to the orbit κ_0, as follows:

$$\Sigma := \{x_0(0) + y \mid y \in \mathbf{R}^{2n} \text{ and } (y, \dot{x}_0(0)) = 0\}. \tag{2.2}$$

In (3.2), $(\cdot, \cdot)$ denotes the scalar product in $\mathbf{R}^{2n}$. Now remark that $R\dot{x}_0(0) = -\dot{x}_0(0)$, and therefore we have for each $y \in \text{Fix}\,(R)$ that

$$(y, \dot{x}_0(0)) = (Ry, \dot{x}_0(0)) = (y, R\dot{x}_0(0)) = -(y, \dot{x}_0(0)) = 0.$$

Also since $x_0(0) \in \text{Fix}\,(R)$, we conclude that Σ contains the n-dimensional subspace Fix (R). In a similar way we define a transversal section Σ_1 to κ_0 at the point $x_0(T_0/2)$ by

$$\Sigma_1 := \left\{ x_0\left(\frac{T_0}{2}\right) + y \;\middle|\; y \in \mathbf{R}^{2n} \text{ and } \left(y, \dot{x}_0\left(\frac{T_0}{2}\right)\right) = 0 \right\}. \tag{2.3}$$

Using (1.6) and the T_0-periodicity of $x_0(t)$ we have $x_0(T_0/2) \in \text{Fix}\,(R)$ and $R\dot{x}_0(T_0/2) = -\dot{x}_0(T_0/2)$. This again implies that Fix $(R) \subset \Sigma_1$.

Now let $\Pi : \tilde{\Sigma} \subset \Sigma \to \Sigma_1$ be the (half-period) Poincaré map between Σ and Σ_1. Π is uniquely defined and smooth in a neighborhood $\tilde{\Sigma}$ of $x_0(0)$ in Σ; moreover, Π is a diffeomorphism from $\tilde{\Sigma}$ onto $\Pi(\tilde{\Sigma})$. Let

$$\Lambda := \text{Fix}\,(R) \cap \Pi(\tilde{\Sigma} \cap \text{Fix}\,(R)).$$

Since $\dim \text{Fix}\,(R) = n$ we have that $\Pi(\tilde{\Sigma} \cap \text{Fix}\,(R))$ is an n-dimensional manifold; moreover, both Fix (R) and $\Pi(\tilde{\Sigma} \cap \text{Fix}\,(R))$ are contained in the $(2n-1)$-dimensional manifold Σ_1. We conclude that generically Λ will be a one-dimensional manifold. Then let $x(t)$ be a solution of (1.1) starting at $t = 0$ from a point $\xi \in \Pi^{-1}(\Lambda)$, and let $T/2$ be the time interval needed to go from ξ to $\Pi(\xi)$; then $x(t)$ satisfies (2.1), and hence $x(t)$ is symmetric and T-periodic. This proves the theorem.

In the foregoing proof we have been rather vague about the precise condition that ensures Λ is a one-dimensional manifold. In § 5 we will give a different analytic proof of Theorem 1, using a precise condition that we now introduce, and that is related to the algebraic structure of the Floquet multiplier 1 of the periodic solution $x_0(t)$ of (1.1). The variational equation of x_0 as a solution of (1.1) is given by

$$\dot{x} = A(t)x, \tag{2.4}$$

with

$$A(t) := Df(x_0(t)) \quad \forall t \in \mathbf{R}. \tag{2.5}$$

Denote by $\Phi(t)$ the fundamental matrix solution of (2.4), satisfying $\Phi(0) = I$. Since (2.4) is T_0-periodic, we have that

$$\Phi(t + T_0) = \Phi(t)C, \qquad C := \Phi(T_0). \tag{2.6}$$

The eigenvalues of the monodromy operator C are the Floquet multipliers of x_0. Since the solutions of (2.4) have the form $x(t) = \Phi(t)\xi$ for some $\xi \in \mathbf{R}^{2n}$ such solution will be T_0-periodic if and only if $C\xi = \xi$, i.e., if and only if ξ is an eigenvector of C corresponding to the multiplier 1. Since $\dot{x}_0(t)$ is a T_0-periodic solution of (2.4) it follows that $\dot{x}_0(t) = \Phi(t)\dot{x}_0(0)$ and that $\dot{x}_0(0)$ is an eigenvector of C corresponding to the multiplier 1.

Let us now denote by $x_\alpha(t)$ ($\alpha \in \mathbf{R}$) a one-parameter family of symmetric, T_α-periodic solutions of (1.1), as given by Theorem 1, coinciding with $x_0(t)$ for $\alpha = 0$ and chosen such that $Rx_\alpha(-t) = x_\alpha(t)$ for all t and α. Suppose also that $x_\alpha(t)$ and T_α depend smoothly on α. We set

$$\omega^*(\alpha) := \frac{T_0}{T_\alpha} \quad \text{and} \quad \beta := \frac{d\omega^*}{d\alpha}(0). \tag{2.7}$$

We have $\omega^*(0) = 1$, and generically we expect that $\beta \neq 0$. Assuming that this is the case we also set

$$x_\alpha^*(t) := x_\alpha\left(\frac{t}{\omega^*(\alpha)}\right) \quad \text{and} \quad u_0(t) := \frac{1}{\beta}\frac{\partial}{\partial \alpha}\bigg|_{\alpha=0} x_\alpha^*(t). \tag{2.8}$$

Then $x_\alpha^*(t)$ and $u_0(t)$ are T_0-periodic, and we have

$$\omega^*(\alpha)\dot{x}_\alpha^*(t) = f(x_\alpha^*(t)). \tag{2.9}$$

Differentiating this identity in the variable α at $\alpha = 0$, we find

$$\dot{u}_0(t) = A(t)u_0(t) - \dot{x}_0(t). \tag{2.10}$$

By $\dot{x}_0(t) = \Phi(t)\dot{x}_0(0)$ and the variation-of-constants formula it follows that

$$u_0(t) = \Phi(t)[u_0(0) - t\dot{x}_0(0)] \quad \forall t \in \mathbf{R}; \tag{2.11}$$

setting $t = T_0$ in (2.11) and using the T_0-periodicity of $u_0(t)$ we find

$$(C - I)u_0(0) = T_0\dot{x}_0(0). \tag{2.12}$$

This proves that (under the condition $\beta \neq 0$) the multiplier 1 of $x_0(t)$ is non-semisimple.

In the next section we will obtain some further results from Floquet theory; in particular we will show that the symmetry (1.6) of $x_0(t)$ implies that the multiplier one of $x_0(t)$ has an even algebraic multiplicity as an eigenvalue of C. Assuming the simplest possible situation that combines even multiplicity with the non-semisimplicity suggested by Theorem 1 and the calculations above we therefore impose the following hypothesis:

(H1) (ii) The Floquet multiplier 1 of $x_0(t)$ is non-semisimple, with geometric multiplicity 1 and algebraic multiplicity 2.

In § 5 we will show that (H0) and (H1) imply the existence of a family of periodic solutions $\{x_\alpha(t)\}$ of (1.1) with the properties assumed above.

3. Floquet theory. Let us recall our problem, as stated in § 1: under hypotheses (H0), (H1), and for given $q \geqq 1$, we want to find all periodic solutions of (1.1) with period near qT_0 and with an orbit near the orbit κ_0 of $x_0(t)$. Using a time rescale, with an undetermined scaling factor $\omega > 0$ for the moment, we can reformulate this problem as follows:

(P_q) Given $q \geqq 1$, find, for all ω near 1, all qT_0-periodic solutions of

$$\omega\dot{x} = f(x) \tag{3.1}$$

that have an orbit in the neighborhood of κ_0.

In this section we fix some (general) $q \geqq 1$; in later sections we will then specialize to the cases $q = 1$, $q = 2$, and $q \geqq 3$, respectively.

To write down (P_q) as an abstract operator equation, we introduce the following Banach spaces:

$$Z_q := \{z: \mathbf{R} \to \mathbf{R}^{2n} \mid z \text{ is continuous and } qT_0\text{-periodic}\}, \tag{3.2}$$

$$X_q := \{x \in Z_q \mid x \text{ is } C^1\}; \tag{3.3}$$

in Z_q we use the C^0-supremum norm, and in X_0 the C^1-supremum norm. We remark that $x_0 \in X_q$. Next we define an operator $M_q: X_q \times \mathbf{R} \to Z_q$ by

$$M_q(x, \omega)(t) := -\omega \dot{x}(t) + f(x(t)) \quad \forall t \in \mathbf{R}. \tag{3.4}$$

Our problem then reduces to that of solving

$$M_q(x, \omega) = 0. \tag{3.5}$$

A fundamental point for our further treatment is the fact that this equation has an $O(2)$-equivariance, as we now explain.

We write $O(2)$, the group of orthogonal linear operators on $\mathbf{R}^2$, in the following form:

$$O(2) = \{\phi \mid \phi \in S^1\} \cup \{\phi \cdot \sigma \mid \phi \in S^1\}, \tag{3.6}$$

where we take $S^1 := \mathbf{R}/q\mathbf{Z}$; ϕ stands for a rotation over $2\pi\phi/q$, and σ stands for a (fixed) reflection. Then we can define representations $\Gamma: O(2) \to \mathscr{L}(Z_q)$ and $\tilde{\Gamma}: O(2) \to \mathscr{L}(Z_q)$ of $O(2)$ over the Banach space Z_q by

$$(\Gamma_\phi z)(t) := z(t + \phi T_0), \qquad (\Gamma_\sigma z)(t) := Rz(-t), \tag{3.7}$$

$$(\tilde{\Gamma}_\phi z)(t) := z(t + \phi T_0), \qquad (\tilde{\Gamma}_\sigma z)(t) := -Rz(-t), \tag{3.8}$$

for each $z \in Z_q$, $\phi \in S^1$, and $t \in \mathbf{R}$. We verify without difficulty that $\Gamma_{\phi_1} \circ \Gamma_{\phi_2} = \Gamma_{\phi_1 + \phi_2}$, $\Gamma_\sigma^2 = I$, and $\Gamma_\sigma \circ \Gamma_\phi = \Gamma_{-\phi} \circ \Gamma_\sigma$, such that Γ is indeed a representation of $O(2)$; the same holds for $\tilde{\Gamma}$. Note also the different sign in the representations Γ_σ and $\tilde{\Gamma}_\sigma$ of σ. The important point is now that the representation Γ leaves the subspace X_q of Z_q invariant and that the operator M_q defined by (3.4) satisfies

$$M_q(\Gamma_\gamma x, \omega) = \tilde{\Gamma}_\gamma M_q(x, \omega) \quad \forall \gamma \in O(2), \tag{3.9}$$

i.e.,

$$M_q(\Gamma_\phi x, \omega) = \tilde{\Gamma}_\phi M_q(x, \omega) \quad \forall \phi \in S^1, \tag{3.10}$$

$$M_q(\Gamma_\sigma x, \omega) = \tilde{\Gamma}_\sigma M_q(x, \omega). \tag{3.11}$$

The fact that M_q is equivariant with respect to two different representations of $O(2)$ is somewhat unusual in equivariant bifurcation theory, but will play a crucial role in our further analysis.

Our starting point will be the given solution $(x, \omega) = (x_0, 1) \in X_q \times \mathbf{R}$ of (3.5). The symmetry of this solution is described by the *isotropy subgroup* of x_0, i.e., by

$$\Sigma_0 := \{\gamma \in O(2) \mid \Gamma_\gamma x_0 = x_0\}. \tag{3.12}$$

It is easily seen that Σ_0 is generated by $\phi = 1$ and by σ: invariance under Γ_1 corresponds to the fact that x_0, considered here as a qT_0-periodic function, is in fact T_0-periodic, whereas the invariance under Γ_σ is a consequence of (1.6). So Σ_0 is isomorphic to D_q, the symmetry group of a regular q-gone.

Since for $\omega = 1$ the solution x_0 of (3.5) does not have the full $O(2)$-symmetry, it generates a whole group orbit of solutions, given by

$$\kappa := \{\Gamma_\gamma \cdot x_0 \mid \gamma \in O(2)\} = \{\Gamma_\phi \cdot x_0 \mid \phi \in S^1\}. \tag{3.13}$$

This group orbit κ generated by x_0 is the counterpart in the space X_q of the actual orbit κ_0 of x_0 in the phase space $\mathbf{R}^{2n}$. We can now reformulate our problem as follows:

(P_q) Find all solutions $(x, \omega) \in X_q \times \mathbf{R}$ of the equation (3.5) with x near κ and ω near 1.

In the next section we will describe an appropriate neighborhood of κ in X_q; in the remainder of this section we study the linearization

$$L_q := D_x M_q(x_0, 1) \in \mathcal{L}(X_q, Z_q) \tag{3.14}$$

of M_q at the solution $(x_0, 1)$. The explicit form of L_q is given by

$$(L_q x)(t) = -\dot{x}(t) + A(t)x(t) \quad \forall x \in X_q, \quad \forall t \in \mathbf{R}, \tag{3.15}$$

while the equivariance of M_q implies that

$$L_q \Gamma_1 = \tilde{\Gamma}_1 L_q \quad \text{and} \quad L_q \Gamma_\sigma = \tilde{\Gamma}_\sigma L_q. \tag{3.16}$$

It is a classical result that $L_q \in \mathcal{L}(X_q, Z_q)$ is a Fredholm operator with index zero. The elements in its kernel $N(L_q)$ are the qT_0-periodic solutions of the variational equation (2.4). Using (2.6), it follows immediately that

$$N(L_q) = \{u(t) = \Phi(t)\xi \mid \xi \in M(C^q - I)\}. \tag{3.17}$$

To describe the range $R(L_q)$ we introduce in Z_q (or, better, in its complexification Z_q^c) an inner product by

$$\langle z, w \rangle := \frac{1}{qT_0} \int_0^{qT_0} (z(t), w(t))\, dt, \qquad (a, b) := \sum_{i=1}^{2n} \bar{a}_i b_i. \tag{3.18}$$

With respect to this inner product L_q has a formal adjoint $L_q^* \in \mathcal{L}(X_q, Z_q)$ defined by

$$(L_q^* x)(t) := \dot{x}(t) + A^T(t)x(t) \quad \forall x \in X_q, \quad \forall t \in \mathbf{R}. \tag{3.19}$$

We have

$$\langle L_q^* x, y \rangle = \langle x, L_q y \rangle \quad \forall x, y \in X_q. \tag{3.20}$$

We can also easily verify that

$$\langle \Gamma_\phi z, w \rangle = \langle z, \Gamma_{-\phi} w \rangle \quad \text{and} \quad \langle \Gamma_\sigma z, w \rangle = \langle z, \Gamma_\sigma w \rangle, \tag{3.21}$$

with a similar property for $\tilde{\Gamma}_\phi$ and $\tilde{\Gamma}_\sigma$. Combining (3.20) and (3.21) with (3.16), we find that

$$L_q^* \tilde{\Gamma}_1 = \Gamma_1 L_q^* \quad \text{and} \quad L_q^* \tilde{\Gamma}_\sigma = \Gamma_\sigma L_q^*. \tag{3.22}$$

Finally, we know from classical Floquet theory that $\dim N(L_q) = \dim N(L_q^*)$,

$$N(L_q^*) = \{u^*(t) = (\Phi^T(t))^{-1}\xi^* \mid \xi^* \in N((C^T)^q - I)\} \tag{3.23}$$

$$R(L_q) = \{z \in Z_q \mid \langle u^*, z \rangle = 0 \quad \forall u^* \in N(L_q^*)\}. \tag{3.24}$$

In what follows we will rewrite $N(L_q)$ and $R(L_q)$ as a direct sum of eigenspaces corresponding to the Floquet multipliers of $x_0(t)$ that are qth roots of unity, i.e., multipliers of the form λ_q^p with $1 \leqq p \leqq q$ and

$$\lambda_q := \exp(2\pi i/q). \tag{3.25}$$

To each $p \in \mathbf{N}$ with $1 \leqq p \leqq q$ we associate a subspace $Z_{(p,q)}$ of Z_q^c defined by

$$Z_{(p,q)} := \{z \in Z_q^c \mid \Gamma_1 z = \lambda_q^p z\}; \tag{3.26}$$

we set $X_{(p,q)} := Z_{(p,q)} \cap X_q^c$. Each of these subspaces $Z_{(p,q)}$ $(1 \leqq p \leqq q)$ coincides with the range of a projection $B_{(p,q)} \in \mathscr{L}(Z_q^c)$ defined by

$$B_{(p,q)} := \frac{1}{q} \sum_{l=1}^{q} \bar{\lambda}_q^{lp} \Gamma_l = \frac{1}{q} \sum_{l=1}^{q} \bar{\lambda}^{lp} \tilde{\Gamma}_l. \tag{3.27}$$

These projections satisfy

$$B_{(p,q)} \circ B_{(p',q)} = \delta_{p,p'} B_{(p,q)}, \quad 1 \leqq p, p' \leqq q, \tag{3.28}$$

$$\sum_{p=1}^{q} B_{(p,q)} = I_{Z_q^c}, \tag{3.29}$$

and

$$\langle B_{(p,q)} z, w \rangle = \langle z, B_{(p,q)} w \rangle. \tag{3.30}$$

It follows that

$$Z_q^c = \bigoplus_{p=1}^{q} Z_{(p,q)} \tag{3.31}$$

and that the spaces $Z_{(p,q)}$ $(1 \leqq p \leqq q)$ are mutually orthogonal:

$$\langle z, w \rangle = 0 \quad \text{if } z \in Z_{(p,q)}, \quad w \in Z_{(p',q)}, \qquad p \neq p' (\mathrm{mod}\, q). \tag{3.32}$$

From (3.16) and (3.22) we see that L_q^c and $(L_q^*)^c$ map each of the spaces $X_{(p,q)}$ into the corresponding space $Z_{(p,q)}$; we define $L_{(p,q)} \in \mathscr{L}(X_{(p,q)}, Z_{(p,q)})$ and $L_{(p,q)}^* \in \mathscr{L}(X_{(p,q)}, Z_{(p,q)})$ as the restrictions of L_q^c, respectively, $(L_q^*)^c$, to $X_{(p,q)}$. It is then immediate from the foregoing that

$$N(L_q^c) = \bigoplus_{p=1}^{q} N(L_{(p,q)}), \qquad N((L_q^*)^c) = \bigoplus_{p=1}^{q} N(L_{(p,q)}^*), \tag{3.33}$$

$$R(L_q^c) = \bigoplus_{p=1}^{q} R(L_{(p,q)}), \tag{3.34}$$

$$R(L_{(p,q)}) = \{ z \in Z_{(p,q)} \mid \langle u^*, z \rangle = 0, \ \forall u^* \in N(L_{(p,q)}^*) \}. \tag{3.35}$$

Using the definitions and (2.6), it is also easy to see that

$$N(L_{(p,q)}) = \{ u(t) = \Phi(t)\xi \mid \xi \in N(C - \lambda_q^p I) \}, \tag{3.36}$$

$$N(L_{(p,q)}^*) = \{ u^*(t) = (\Phi^T(t))^{-1} \xi^* \mid \xi^* \in N(C^T - \bar{\lambda}_q^p I) \}. \tag{3.37}$$

Next we prove a general result on Floquet multipliers of periodic solutions of (1.1). Let $\tilde{x}(t)$ be a periodic solution of (1.1), with period $T > 0$. Let $\tilde{\Phi}(t)$ be the fundamental matrix solution of the variational equation

$$\dot{x} = Df(\tilde{x}(t))x, \tag{3.38}$$

and $\tilde{C} := \tilde{\Phi}(T)$ the corresponding monodromy operator. Let $\omega := T_0/T$, and define $\tilde{x}_\omega : \mathbf{R} \to \mathbf{R}^{2n}$ by $\tilde{x}(t) = \tilde{x}_\omega(\omega t)$; then $\tilde{x}_\omega$ is T_0-periodic, and hence $\tilde{x}_\omega \in X_q$ for each $q \geqq 1$, and the operator $D_x M_q(\tilde{x}_\omega, \omega) \in \mathscr{L}(X_q, Z_q)$ is well defined. Moreover, this operator again maps each subspace $X_{(p,q)}$ $(1 \leqq p \leqq q)$ of X_q^c into the corresponding subspace $Z_{(p,q)}$ of Z_q^c. We define $\tilde{L}_{(p,q)} \in \mathscr{L}(X_{(p,q)}, Z_{(p,q)})$ as the restriction of $D_x M_q(\tilde{x}_\omega, \omega)$ to $X_{(p,q)}$. Finally, we denote by $J_{(p,q)} : X_{(p,q)} \to Z_{(p,q)}$ the canonical injection of $X_{(p,q)}$ into $Z_{(p,q)}$.

LEMMA 1. *With the foregoing notation we have for each $\mu \in \mathbf{C}$ and each $\nu \geqq 1$ that*

$$\dim N((\tilde{L}_{(p,q)} - \mu J_{(p,q)})^\nu) = \dim N((\tilde{C} - \lambda_q^p e^{\mu T} I)^\nu). \tag{3.39}$$

Proof. Fix some $\mu \in \mathbf{C}$ and a pair (p, q) with $1 \leqq p \leqq q$; we have that $\tilde{\lambda} := \lambda_q^p e^{\mu T} = e^{\tilde{\mu} T}$, with $\tilde{\mu} = \omega p \mu_q + \mu$ and $\mu_q := 2\pi i / qT_0$. Since $\det \tilde{C} \neq 0$ it follows from a classical construction (involving the Jordan normal form of $\tilde{C}$) that we can find some $E \in \mathscr{L}(\mathbf{C}^{2n})$ such that

$$\tilde{C} = e^{ET} \quad \text{and} \quad N(E - \tilde{\mu} I) = N(\tilde{C} - \tilde{\lambda} I). \tag{3.40}$$

Now let $u \in X_{(p,q)}$ and $v \in Z_{(p,q)}$ be such that

$$\tilde{L}_{(p,q)} u - \mu u = v. \tag{3.41}$$

Define functions $x : \mathbf{R} \to \mathbf{C}^{2n}$ and $y : \mathbf{R} \to \mathbf{C}^{2n}$ by

$$u(\omega t) = \tilde{\Phi}(t)\, e^{-(E - \omega p \mu_q) t} x(t), \tag{3.42}$$

$$v(\omega t) = \tilde{\Phi}(t)\, e^{-(E - \omega p \mu_q) t} y(t); \tag{3.43}$$

it is then straightforward to check that $x(t)$ and $y(t)$ are T-periodic, and that (3.41) transforms into

$$-\dot{x} + (E - \tilde{\mu} I) x = y. \tag{3.44}$$

Conversely, if $x(t)$ and $y(t)$ are T-periodic and satisfy (3.44), then $u(t)$ and $v(t)$ defined by (3.42) and (3.43) belong to $X_{(p,q)}$, respectively $Z_{(p,q)}$, and (3.41) holds. The left-hand side of (3.44) defines an operator $K \in \mathscr{L}(X_T, Z_T)$, where Z_T is the space of continuous T-periodic mappings $z : \mathbf{R} \to \mathbf{C}^{2n}$, and X_T the subspace of all $x \in Z_T$ which are of class C^1. We will prove that

$$N(K^\nu) = \{\zeta(\cdot) \equiv \xi \mid \xi \in N((\tilde{C} - \tilde{\lambda} I)^\nu)\}, \tag{3.45}$$

for each $\nu \geqq 1$. By the established relation between (3.41) and (3.44) this immediately implies the result (3.39). We prove (3.45) using induction in ν.

For a C^1-function $\zeta : \mathbf{R} \to \mathbf{C}^{2n}$ the condition $\zeta \in N(K)$ is equivalent to

$$\zeta(t) = e^{(E - \tilde{\mu} I) t} \zeta(0) \quad \text{and} \quad \zeta(T) = \zeta(0),$$

i.e., to

$$\zeta(t) = e^{(E - \tilde{\mu} I) t} \zeta(0) \quad \text{and} \quad \zeta(0) \in N(\tilde{C} - \tilde{\lambda} I).$$

Since $N(\tilde{C} - \tilde{\lambda} I) = N(E - \tilde{\mu} I)$ this reduces to $\zeta(t) \equiv \zeta(0)$ with $\zeta(0) \in N(\tilde{C} - \tilde{\lambda} I)$; this proves (3.45) for $\nu = 1$.

Now suppose that (3.45) holds for a certain $\nu \geqq 1$, and let $\zeta : \mathbf{R} \to \mathbf{C}^{2n}$ be a C^1-function. We then have that $\zeta \in M(K^{\nu+1})$ if and only if $\zeta \in X_T$ and $K\zeta \in N(K^\nu)$. By (3.45), the definition of K, and the variation-of-constants formula this is equivalent to the existence of some $\xi \in N((\tilde{C} - \tilde{\lambda} I)^\nu)$ such that

$$\zeta(t) = e^{(E - \tilde{\mu} I) t} \left[\zeta(0) - \int_0^t e^{-(E - \tilde{\mu} I) s} \xi \, ds \right] \quad \text{and} \quad \zeta(T) = \zeta(0). \tag{3.46}$$

For a function $\zeta(t)$ having the form given by (3.46), the condition $\zeta(T) = \zeta(0)$ is equivalent to

$$(\tilde{C} - \tilde{\lambda} I) \zeta(0) = \tilde{C} \int_0^T e^{-(E - \tilde{\mu} I) s} \xi \, ds. \tag{3.47}$$

Writing $\xi = (E - \tilde{\mu} I) \zeta(0) + \xi'$ in (3.47), applying $(E - \tilde{\mu} I)$, and working out the integral gives $(\tilde{C} - \tilde{\lambda} I) \xi' = 0$, i.e., $\xi' \in N(\tilde{C} - \tilde{\lambda} I) = N(E - \tilde{\mu} I)$; but then (3.47) reduces to $C T \xi' = 0$, i.e., $\xi' = 0$ and $\xi = (E - \tilde{\mu} I) \zeta(0)$. Bringing this into (3.47) we see that the condition $\zeta \in N(K^{\nu+1})$ is equivalent to $\zeta(t) \equiv \zeta(0)$ with $(E - \tilde{\mu} I) \zeta(0) \in N((\tilde{C} - \tilde{\lambda} I)^\nu)$; using (3.40) this last condition is equivalent to $\zeta(0) \in N((\tilde{C} - \tilde{\lambda} I)^\nu (E - \tilde{\mu} I)) = N((\tilde{C} - \tilde{\lambda} I)^{\nu+1})$. This proves (3.45) for all $\nu \geqq 1$.

Taking $\nu = 1$ in (3.39) we see that $\tilde{\lambda} = \lambda_q^p e^{\mu T}$ is a Floquet multiplier of $\tilde{x}(t)$ if and only if $\dim N(\tilde{L}_{(p,q)} - \mu J_{(p,q)}) > 0$, i.e., if and only if the equation

$$D_x M_q(\tilde{x}_\omega, \omega) \cdot u = \mu u \tag{3.48}$$

has a nontrivial solution $u \in X_{(p,q)}$. Also, taking $\tilde{x}(t) = x_0(t)$ and $\mu = 0$ it follows that

$$\dim N(L^\nu_{(p,q)}) = \dim N((C - \lambda_q^p I)^\nu) \quad \forall \nu \geqq 1. \tag{3.49}$$

Until now we have not yet exploited the fact that $x_0(t)$ is symmetric with respect to time reversion. To see how this can be used we prove the following complement of Lemma 1.

LEMMA 2. *Suppose that the T-periodic solution $\tilde{x}(t)$ in Lemma 1 is symmetric. Then we have for each $\lambda \in \mathbf{C}$, $\lambda \neq 0$ that*

$$\dim N((\tilde{C} - \lambda I)^\nu) = \dim N((\tilde{C} - \lambda^{-1} I)^\nu) \quad \forall \nu \geqq 1. \tag{3.50}$$

Proof. Since $\tilde{x}(t)$ is symmetric we will have $\Gamma_\phi \Gamma_\sigma \tilde{x}_\omega = \tilde{x}_\omega$ for some $\phi \in \mathbf{R}$. Let $\mu \in \mathbf{C}$ be such that $\lambda = e^{\mu T}$, take $p = q = 1$, and write $\tilde{L}$ and J for $\tilde{L}_{(1,1)}$ and $J_{(1,1)}$, respectively. Now $\tilde{L} = D_x M_1(\tilde{x}_\omega, \omega)^c \in \mathscr{L}(X_1^c, Z_1^c)$, and therefore the equivariance of M_1, the symmetry of $\tilde{x}_\omega$, and $J\Gamma_\phi \Gamma_\sigma = -\tilde{\Gamma}_\phi \tilde{\Gamma}_\sigma J$ imply that

$$(\tilde{L} - \mu J)\Gamma_\phi \Gamma_\sigma = \tilde{\Gamma}_\phi \tilde{\Gamma}_\sigma (\tilde{L} + \mu J) \quad \forall \mu \in \mathbf{C}, \tag{3.51}$$

i.e., we have $\tilde{L} + \mu J = \tilde{\Gamma}_\phi \tilde{\Gamma}_\sigma (\tilde{L}_\mu J)\Gamma_\phi \Gamma_\sigma$. It follows that

$$N((\tilde{L} + \mu J)^\nu) = \Gamma_\phi \Gamma_\sigma (N((\tilde{L} - \mu J)^\nu)),$$

$$\dim N((\tilde{L} + \mu J)^\nu) = \dim N((\tilde{L} - \mu J)^\nu),$$

and (3.50) follows from (3.39).

We now discuss some consequences of Lemma 2. It follows from (3.50) that if $\lambda \in \mathbf{C}$ is a multiplier of $\tilde{x}(t)$, then also λ^{-1} is a multiplier (multipliers are always different from zero), and both have the same algebraic multiplicity. Therefore, we can divide the multipliers into groups, of the following forms:

(1) Pairs of real multipliers $\{\lambda, \lambda^{-1}\}$, with $|\lambda| \neq 1$;

(2) Pairs of complex multipliers $\{\lambda, \bar{\lambda}\}$ on the unit circle (i.e., $|\lambda| = 1$, but $\lambda \neq \pm 1$);

(3) Quadruples of complex multipliers $\{\lambda, \lambda^{-1}, \bar{\lambda}, \bar{\lambda}^{-1}\}$ off the unit circle and off the real axis (i.e., $|\lambda| \neq 1$ and $\operatorname{Im} \lambda \neq 0$);

(4) Possibly the multiplier -1;

(5) The multiplier 1, which is always present.

Counting the dimensions of the corresponding generalized eigenspaces of $\tilde{C}$, and using the fact that $\det \tilde{C} > 0$ (this is true in general), we conclude that -1 must be a multiplier with even multiplicity, that $\det \tilde{C} = 1$, and that, since we are working in an even-dimensional phase space, the multiplier 1 must also have even multiplicity. For generic equations of type (1, 1) and for generic symmetric periodic solutions $\tilde{x}(t)$ the multiplier 1 will be non-semisimple with algebraic multiplicity 2 (for $\tilde{x} = x_0$ we have already anticipated this in hypothesis (H1)(ii)), -1 will not be a multiplier, all other multipliers will be simple, and those on the unit circle will not be roots of unity. However, along one-parameter families of symmetric periodic solutions such as we find in Theorem 1, we may generically find critical parameter values for which we have either (a) a pair of simple multipliers on the unit circle that are roots of unity; or (b) a multiplier -1 that is non-semisimple with algebraic multiplicity 2. The roots of unity appear as simple multipliers move along the unit circle, whereas the multiplier -1 appears at a transition point where two real simple multipliers $\{\lambda, \lambda^{-1}\}$ meet at -1 and then move away from each other on the unit circle. Of course, a similar situation may happen at

1, but here we will not discuss this (more difficult) case. Also, in a quadruple $\{\lambda, \bar{\lambda}, \lambda^{-1}, \bar{\lambda}^{-1}\}$ we may have that λ and $\bar{\lambda}^{-1}$ meet on the unit circle and then separate again, staying now on the unit circle; however, generically the multiplier at the critical value of the parameter will not be a root of unity, and therefore we will not consider that case either. The foregoing discussion will motivate our hypotheses (H2) and (Hq), ($q \geqq 3$), which we introduce later.

4. Reduction of the problem. In this section we make a first step towards the solution of (3.5), which we want to solve for (x, ω) near $\kappa \times \{1\}$ in $X_q \times \mathbf{R}$. We will reduce this problem to that of solving a new equation (4.12), which is Σ_0-equivariant and which we must solve near the origin; in the next sections we will then apply a standard Lyapunov-Schmidt reduction to this reduced equation.

We start by using the Floquet theory of § 3 to obtain some consequences of our hypothesis (H1). Setting $\tilde{x} = x_0$, $\mu = 0$, and $p = q = 1$ in Lemma 1 we see that (H1)(ii) implies that

$$\dim N(L_1) = 1 \quad \text{and} \quad \dim N(L_1^\nu) = 2 \quad \forall \nu \geqq 2. \tag{4.1}$$

Since $\dot{x}_0 \in N(L_1)$, it follows that

$$N(L_1) = \operatorname{span}\{\dot{x}_0\}. \tag{4.2}$$

LEMMA 3. *Assume* (H0)-(H1). *Then there exist uniquely determined elements* $u_0 \in X_1$, $u_0^* \in X_1$ *and* $\tilde{u}_0^* \in X_1$ *satisfying the following properties*:

(i) $L_1 u_0 = \dot{x}_0$ *and* $\langle u_0, \dot{x}_0 \rangle = 0$;

(ii) $N(L_1^*) = \operatorname{span}\{u_0^*\}$ *and* $\langle \bar{u}_0^*, u_0 \rangle = 1$;

(iii) $L_1^* \tilde{u}_0^* = u_0^*$ *and* $\langle \tilde{u}_0^*, u_0 \rangle = 0$.

Moreover, we then have that

$$R(L_1) = \{z \in Z_1 \mid \langle u_0^*, z \rangle = 0\}, \tag{4.3}$$

and the functions u_0, u_0^*, *and* $\tilde{u}_0^*$ *also satisfy*

(iv) $\langle u_0^*, \dot{x}_0 \rangle = 0$ *and* $\langle \tilde{u}_0^*, \dot{x}_0 \rangle = 1$;

(v) $\Gamma_\sigma u_0 = u_0$, $\tilde{\Gamma}_\sigma u_0^* = -u_0^*$, *and* $\tilde{\Gamma}_\sigma \tilde{u}_0^* = \tilde{u}_0^*$.

Proof. By (4.1) there exists some $u \in N(L_1^2) \backslash N(L_1)$. Then $L_1 u \in N(L_1) \backslash \{0\}$, and hence, by (4.2), $L_1 u = \beta \dot{x}_0$ for some $\beta \neq 0$. We can then satisfy (i) by taking

$$u_0 = \beta^{-1}(u - \langle \dot{x}_0, u \rangle \langle \dot{x}_0, \dot{x}_0 \rangle^{-1} \dot{x}_0).$$

The uniqueness of $u_0 \in X_1$ satisfying (i) is easily verified.

Next we observe that $u_0 \notin R(L_1)$; indeed, if $u_0 = L_1 x$ for some $x \in X_1$ then $x \in N(L_1^3) = N(L_1^2)$, and $L_1 u_0 = L_1^2 x = 0$, which contradicts (i). It then follows from (3.24) that there exists some $u^* \in N(L_1^*)$ such that $\langle x^*, u_0 \rangle \neq 0$; moreover, since $\dim N(L_1^*) = \dim N(L_1) = 1$, we also have $N(L_1^*) = \operatorname{span}\{u^*\}$. We then satisfy (ii) by taking $u_0^* = \langle u^*, u_0 \rangle^{-1} u^*$; uniqueness of u_0^* is again trivial.

It follows from (i) that $\dot{x}_0 \in R(L_1)$, and hence $\langle u_0^*, \dot{x}_0 \rangle = 0$. Reversing the roles of L_1 and L_1^* (L_1 is the formal adjoint of L_1^*) we see that this implies that $u_0^* \in R(L_1^*)$. So we can find some $\tilde{u}^* \in X_1$ such that $L_1^* \tilde{u}^* = u_0^*$. Setting $\tilde{u}_0^* = \tilde{u}^* - \langle \tilde{u}^*, u_0 \rangle u_0^*$, we then satisfy (iii); also here uniqueness is easy. Moreover, we have that

$$\langle \tilde{u}_0^*, \dot{x}_0 \rangle = \langle \tilde{u}_0^*, L_1 u_0 \rangle = \langle L_1^* \tilde{u}_0^*, u_0 \rangle = \langle u_0^*, u_0 \rangle = 1.$$

To prove (v) we use (3.16) and (3.22) to show that properties (i)-(iii) remain satisfied when we replace u_0 by $\Gamma_\sigma u_0$, u_0^* by $-\tilde{\Gamma}_\sigma u_0^*$, and $\tilde{u}_0^*$ by $\tilde{\Gamma}_\sigma \tilde{u}_0^*$; the uniqueness of the elements u_0, u_0^*, and $\tilde{u}_0^*$ satisfying (i)-(iii) then implies (v).

Remark. Further on we will consider u_0, u_0^*, and $\tilde{u}_0^*$ as elements of X_q, for some $q \geqq 1$. As such they do not only have the symmetry given by (v), but also

$$\Gamma_1 u_0 = u_0, \quad \tilde{\Gamma}_1 u_0^* = u_0^*, \quad \tilde{\Gamma}_1 \tilde{u}_0^* = u_0^*. \tag{4.4}$$

Next we consider the problem of describing a neighborhood of κ in X_q. As a solution of a differential equation with smooth right-hand side, $x_0(t)$ is smoother than a general element of X_q (which is only C^1); in particular, the mapping $\phi \mapsto \Gamma_\phi x_0$ is differentiable, and its derivative determines the tangent space to κ at x_0. Since this derivative equals $T_0 \dot{x}_0$, we see that this tangent space is generated by $\dot{x}_0$. We now introduce a projection P_0 in Z_q onto this tangent space, by setting

$$P_0 z := \langle \tilde{u}_0^*, z \rangle \dot{x}_0. \tag{4.5}$$

It follows from (4.4) and Lemma 3(v) that

$$P_0 \Gamma_1 = \Gamma_1 P_0 = P_0 \quad \text{and} \quad P_0 \Gamma_\sigma = \Gamma_\sigma P_0 = -P_0. \tag{4.6}$$

Therefore, if we set

$$Y_q := N(P_0) = \{ z \in Z_q \mid \langle \tilde{u}_0^*, z \rangle = 0 \}, \tag{4.7}$$

then Y_q is invariant under the action of $\Sigma_0 = D_q$.

The following lemma now describes a tubular neighborhood of κ in X_q; for its proof we refer to Chapter 8 of [9].

LEMMA 4. *There exists a Σ_0-invariant neighborhood Ω of the origin in $X_q \cap Y_q$ such that*

$$\{ \Gamma_\gamma \cdot (x_0 + y) \mid y \in \Omega, \gamma \in O(2) \}$$

forms an open neighborhood of κ in X_q.

This result allows us to write $x = \Gamma_\gamma \cdot (x_0 + y)$ in (3.5); using the equivariance of M_q and the invertibility of $\tilde{\Gamma}_\gamma$ we then arrive at

$$M_q(x_0 + y, \omega) = 0. \tag{4.8}$$

This means that by solving (4.8) for (y, ω) near $(0, 1)$ in $(X_q \cap Y_q) \times \mathbf{R}$ and acting with the symmetry operators on the solution set we obtain all solutions (x, ω) of (4.5) in a neighborhood of $\kappa \times \{1\}$ in $X_q \times \mathbf{R}$. We remark that in (4.8) y belongs to a subspace of codimension 1, namely, $X_q \cap Y_q = \{ y \in X_q \mid \langle \tilde{x}_0^*, y \rangle = 0 \}$.

For our next step we rewrite (4.8) as two separate equations:

$$P_0 M_q(x_0 + y, w) = 0, \tag{4.9}$$

$$(I - P_0) M_q(x_0 + y, \omega) = 0. \tag{4.10}$$

Now we observe that the first of these equations can be solved for ω. Indeed, (4.9) is satisfied for $(y, \omega) = (0, 1)$, and using the definitions (3.4) and (4.5) of M_q and P_0, respectively, we find that

$$P_0 D_\omega M_q(x_0, 1) = -\dot{x}_0 \neq 0.$$

So we can solve (4.9) by the Implicit Function Theorem, to obtain $\omega = \tilde{\omega}(y)$, where $\tilde{\omega} : X_q \cap Y_q \to \mathbf{R}$ is defined and smooth in a neighborhood of the origin, with $\tilde{\omega}(0) = 1$. Moreover, the uniqueness part of the Implicit Function Theorem and the equivariance properties of M_q and P_0 imply that

$$\tilde{\omega}(\Gamma_\gamma y) = \tilde{\omega}(y) \quad \forall \gamma \in \Sigma_0 = D_q. \tag{4.11}$$

Bringing the solution $\omega = \tilde{\omega}(y)$ of (4.9) into (4.10) gives us the reduced equation

$$\tilde{M}_q(y) := (I - P_0) M_q(x_0 + y, \tilde{\omega}(y)) = 0, \tag{4.12}$$

to be solved for y near the origin in $X_q \cap Y_q$. The mapping $\tilde{M}_q : X_q \cap Y_q \to Y_q$ is defined and smooth near the origin, $\tilde{M}_q(0) = 0$, and $\tilde{M}_q$ is D_q-equivariant:

$$\tilde{M}_q(\Gamma_\gamma \cdot y) = \tilde{\Gamma}_\gamma M_q(y) \quad \forall \gamma \in D_q. \tag{4.13}$$

For later use we also calculate $\tilde{L}_q := D\tilde{M}_q(0)$, $D^2\tilde{M}_q(0)$, and $D\tilde{\omega}(0)$. We have from (4.12) that

$$\begin{aligned} \tilde{L}_q y &= (I - P_0)L_q y + (I - P_0)D_\omega M_q(x_0, 1) \cdot D\tilde{\omega}(0) \cdot y \\ &= (I - P_0)L_q y \quad \forall y \in X_q \cap Y_q, \end{aligned} \tag{4.14}$$

since $D_\omega M_q(x_0, 1) = -\dot{x}_0$ and $(I - P_0)\dot{x}_0 = 0$. To find $D\tilde{\omega}(0)$ we differentiate the identity

$$P_0 M_q(x_0 + y, \tilde{\omega}(y)) = 0$$

at $y = 0$, and find

$$D\tilde{\omega}(0) \cdot y = \langle \tilde{u}_0^*, L_q y \rangle = \langle L_q^* \tilde{u}_0^*, y \rangle = \langle u_0^*, y \rangle. \tag{4.15}$$

Finally, differentiating (4.12) twice and setting $y = 0$ gives

$$\begin{aligned} D^2\tilde{M}_q(0) \cdot (y, \tilde{y}) &= (I - P_0)D_x^2 M_q(x_0, 1) \cdot (y, \tilde{y}) \\ &\quad - (D\tilde{\omega}(0) \cdot \tilde{y})(I - P_0)\dot{y} - (D\tilde{\omega}(0) \cdot y)(I - P_0) \cdot \dot{\tilde{y}} \\ &= (I - P_0)D_x^2 M_q(x_0, 1) \cdot (y, \tilde{y}) \\ &\quad - \langle u_0^*, \tilde{y} \rangle (I - P_0)\dot{y} - \langle u_0^*, y \rangle (I - P_0)\dot{\tilde{y}}. \end{aligned} \tag{4.16}$$

Remark. The foregoing reduction of (3.5) to (4.12) is a kind of analogue of the classical Poincaré map construction, but here in an infinite-dimensional space of periodic mappings. As is the case with the Poincaré map we have reduced the problem to one on a subspace of codimension 1, and we have gotten rid of ω, which is related to the period. The important point here is that the reduced equation (4.12) reflects the symmetry of our basic solution x_0, since it is Σ_0-equivariant.

The next step toward a solution of our problem will be to apply a classical (equivariant) Lyapunov-Schmidt reduction to (4.30); we will do this in the sections that follow. Here we conclude this section with a result on the operator $\tilde{L}_q \in \mathcal{L}(X_q \cap Y_q, Y_q)$ given by (4.14).

LEMMA 5. *Assume* (H0)-(H1), *and* $q \geqq 1$. *Then* $\tilde{L}_q \in \mathcal{L}(X_q \cap Y_q, Y_q)$ *is a Fredholm operator with index zero,* $\dim N(\tilde{L}_q) = \dim N(L_q)$,

$$N(\tilde{L}_q) = (N(L_q) \cap Y_q) \oplus \operatorname{span}\{u_0\}, \tag{4.17}$$

$$R(\tilde{L}_q) = R(L_q) \cap Y_q. \tag{4.18}$$

Proof. Let $u \in N(\tilde{L}_q)$; by (4.14) this means that $u \in X_q \cap Y_q$ and $(I - P_0)L_q u = 0$, i.e., $L_q u = \alpha \dot{x}_0$ for some $\alpha \in \mathbf{R}$. Let $v := u - \alpha u_0$; since $u_0 \in Y_q$ it follows that also $v \in Y_q$, while $L_q v = 0$, i.e., we have $v \in N(L_q) \cap Y_q$ and therefore $u \in (N(L_q) \cap Y_q) \oplus \operatorname{span}\{u_0\}$. Conversely, let $u = \alpha u_0 + v$, with $\alpha \in \mathbf{R}$ and $v \in N(L_q) \cap Y_q$. Then $u \in X_q \cap Y_q$ and $\tilde{L}_q u = 0$. This proves (4.17).

Next let $z \in R(\tilde{L}_q)$, i.e., there exists some $y \in X_q \cap Y_q$ such that $z = \tilde{L}_q y = (I - P_0)L_0 y$; it follows that $z \in Y_q$ and $L_q y = z + \beta \dot{x}_0$ for some $\beta \in \mathbf{R}$. But then $L_q(y - \beta u_0) = z$, which proves that $z \in R(L_q) \cap Y_q$. Conversely, if $z \in R(L_q) \cap Y_q$, then $z = L_q x$ for some $x \in X_q$. Set $y = (I - P_0)x$; then $y \in X_q \cap Y_q$, $L_q y = z$, and $\tilde{L}_q y = (I - P_0)L_q y = (I - P_0)z = z$, since $z \in Y_q$. This shows that $z \in R(\tilde{L}_q)$ and proves (4.18).

Now $L_q \in \mathscr{L}(X_q, Z_q)$ is a Fredholm operator with index zero. Since $\dot{x}_0 \in N(L_q)$ and $u_0 \notin N(L_q)$, it follows from (4.17) that $\dim N(\tilde{L}_q) = \dim N(L_q)$. It is obvious from (4.18) that $R(\tilde{L}_q)$ is closed. To calculate the codimension of $R(\tilde{L}_q)$ in Y_q, we observe that (3.24) and (4.18) imply

$$\begin{aligned} R(\tilde{L}_q) &= \{y \in Y_q \mid \langle u^*, y\rangle = 0,\ \forall u^* \in N(L_q^*)\} \\ &= \{y \in Z_q \mid \langle \tilde{u}_0^*, y\rangle = 0 = \langle u^*, y\rangle,\ \forall u^* \in N(L_q^*)\}. \end{aligned} \tag{4.19}$$

Since $\tilde{u}_0^* \notin N(L_q^*)$, it follows that $R(\tilde{R}_q)$ has codimension equal to $1 + \dim N(L_q^*)$ in Z_q, and equal to $\dim N(L_q^*) = \dim N(L_q) = \dim N(\tilde{L}_q)$ in Y_q. This proves the lemma.

Using the results of § 3, and in particular (3.31)-(3.34), we can give some more detailed expressions for $N(\tilde{L}_q)$ and $R(\tilde{L}_q)$. Indeed, it follows from (3.32) and $\tilde{u}_0^* \in X_1 \subset Z_1^c = Z_{(q,q)}$ that $Z_{(p,q)} \subset Y_q$ if $1 \leqq p \leqq q-1$. It follows then from (3.33), (4.17), and $N(L_{(q,q)}) = N(L_1^c) = \operatorname{span}\{\dot{x}_0\}$ that

$$N(\tilde{L}_q) = \operatorname{span}\{u_0\} \oplus \operatorname{Re}\left(\bigoplus_{p=1}^{q-1} N(L_{(p,q)})\right); \tag{4.20}$$

the second term on the right-hand side of (4.20) is absent if $q = 1$. A similar argument shows that (for $q \geqq 2$)

$$R(\tilde{L}_q) = R(\tilde{L}_1) \oplus \operatorname{Re}\left(\bigoplus_{p=1}^{q-1} R(L_{(p,q)})\right). \tag{4.21}$$

Setting $q = 1$ in (4.19) and using $N(L_1^*) = \operatorname{span}\{u_0^*\}$ finally gives us

$$R(\tilde{L}_1) = \{y \in Y_1 \mid \langle u_0^*, y\rangle = 0\}. \tag{4.22}$$

5. The case $q = 1$. In this section we consider (4.12) in the case where $q = 1$. Our main result is the following theorem, which in fact is a more precise version of our earlier Theorem 1.

THEOREM 2. *Assume* (H0) *and* (H1). *Then there exist a neighborhood $\mathscr{U}$ of $\kappa \times \{1\}$ in $X_1 \times \mathbf{R}$, a number $\alpha_0 > 0$, and smooth mappings $x^*:]-\alpha_0, \alpha_0[\to X_1$ and $\omega^*:]-\alpha_0, \alpha_0[\to \mathbf{R}$ such that the following hold*:

(i) $\{(x, \omega) \in \mathscr{U} \mid M_1(x, \omega) = 0\} = \{(\Gamma_\phi \cdot x^*(\alpha), \omega^*(\alpha)) \mid |\alpha| < \alpha_0, \phi \in S^1\}$;

(ii) $(x^*(0), \omega^*(0)) = (x_0, 1)$;

(iii) $\langle u_0^*, x^*(\alpha) - x_0\rangle = \alpha$ *and* $\langle \tilde{u}_0^*, x^*(\alpha) - x_0\rangle = 0$;

(iv) $\Gamma_\sigma x^*(\alpha) = x^*(\alpha)$.

Proof. We solve (4.12) for $q = 1$ and y near the origin. It follows from (4.20) and (4.22) that

$$N(\tilde{L}_1) = \operatorname{span}\{u_0\}, \qquad R(\tilde{L}_1) = \{y \in Y_1 \mid \langle u_0^*, y\rangle = 0\}. \tag{5.1}$$

We define a projection P in Y_1 by

$$Py := \langle u_0^*, y\rangle u_0. \tag{5.2}$$

Then $N(\tilde{L}_1) = R(P)$ and $R(\tilde{L}_1) = N(P)$. We also remark that

$$P\Gamma_\sigma = \Gamma_\sigma P. \tag{5.3}$$

We can now apply a standard Lyapunov-Schmidt reduction to (4.12). We write $y \in X_1 \cap Y_1$ as $y = \alpha u_0 + v$, with $v \in N(P)$, and rewrite (4.12) in the following form:

$$(I - P)\tilde{M}_1(\alpha u_0 + v) = 0, \tag{5.4a}$$

$$P\tilde{M}_1(\alpha u_0 + v) = 0. \tag{5.4b}$$

Equation (5.4a) can be solved for $v = v^*(\alpha)$, with $v^*(0) = 0$ and $\Gamma_\sigma v^*(\alpha) = v^*(\alpha)$ (since $\Gamma_\sigma u_0 = u_0$). Bringing this solution into (5.4b) and using definition (5.2) of P leaves us with a scalar bifurcation equation:

$$g_0(\alpha) := \langle u_0^*, \tilde{M}_1(\alpha u_0 + v^*(\alpha))\rangle = 0. \tag{5.5}$$

But

$$\begin{aligned} g_0(\alpha) &= \langle u_0^*, \tilde{M}_1(\Gamma_\sigma(\alpha u_0 + v^*(\alpha)))\rangle = \langle u_0^*, \tilde{\Gamma}_\sigma \tilde{M}_1(\alpha u_0 + v^*(\alpha))\rangle \\ &= \langle \tilde{\Gamma}_\sigma u_0^*, \tilde{M}_1(\alpha u_0 + v^*(\alpha))\rangle = -g_0(\alpha), \end{aligned}$$

and hence $g_0(\alpha) \equiv 0$. Putting $x^*(\alpha) := x_0 + \alpha u_0 + v^*(\alpha)$, $\omega^*(\alpha) := \tilde{\omega}(\alpha u_0 + v^*(\alpha))$, and using the results of § 4 then proves the theorem.

Differentiating the identity $(I - P)\tilde{M}_1(\alpha u_0 + v^*(\alpha)) = 0$, we find that $Dv^*(0) = 0$, and hence

$$Dx^*(0) = u_0. \tag{5.6}$$

Using (4.15) we find that

$$D\omega^*(0) = \langle u_0^*, u_0\rangle = 1. \tag{5.7}$$

It follows from Theorem 2 that when we put

$$x_\alpha(t) := x^*(\alpha)(\omega^*(\alpha)t), \tag{5.8}$$

then $\{x_\alpha \,|\, |\alpha| < \alpha_0\}$ forms a solution branch of our original equation, consisting of symmetric periodic solutions with period

$$T_\alpha := \frac{T_0}{\omega^*(\alpha)}. \tag{5.9}$$

It follows from (5.7) that along this solution branch the period changes with nonzero speed. Using the arguments of § 2 we conclude that for $|\alpha|$ sufficiently small the solution $x_\alpha(t)$ of (1.1) will have a non-semisimple Floquet multiplier 1, with geometric multiplicity 2.

When we move along the one-parameter family of periodic solutions $\{x_\alpha\}$ we may (as we have discussed in § 4) encounter in a generic way critical values of the parameter α for which $x_\alpha(t)$ has either (a) a multiplier -1, which is non-semisimple and has algebraic multiplicity 2; or (b) a pair of simple Floquet multipliers on the unit circle that are roots of unity. In the sections that follow we will consider separately the cases (a) and (b), assuming that the critical parameter value is given by $\alpha = 0$.

6. The case $q = 2$. In this section we study (4.12) for the case $q = 2$, and assuming that next to (H0) and (H1) we also have:

(H2) (i) $x_0(t)$ has -1 as a non-semisimple Floquet multiplier, with geometric multiplicity 1 and algebraic multiplicity 2.

We will state part (ii) of (H2) after we have studied what happens to this multiplier -1 as we move along the branch of periodic solutions $\{x_\alpha\}$ given by § 5.

We start with some notation. We set $Z_2^- := \mathrm{Re}\,(Z_{(1,2)}) = \{z \in Z_2 \,|\, \Gamma_1 z = -z\}$ and $X_2^- := \mathrm{Re}\,(X_{(1,2)})$; we denote by $L_2^- \in \mathscr{L}(X_2^-, Z_2^-)$ and $(L_2^*)^- \in \mathscr{L}(X_2^-, Z_2^-)$ the restrictions of L_2 and L_2^*, respectively, to X_2^-. It then follows immediately from (4.20), (4.21), and (3.35) that

$$N(\tilde{L}_2) = \mathrm{span}\,\{u_0\} \oplus N(L_2^-), \tag{6.1}$$

$$R(\tilde{L}_2) = R(\tilde{L}_1) \oplus R(L_2^-), \tag{6.2}$$

$$R(L_2^-) = \{z \in Z_2^- \,|\, \langle x^*, z\rangle = 0,\ \forall u^* \in N((L_2^-)^*)\}. \tag{6.3}$$

Moreover, taking $\tilde{x} = x_0$, $p = 1$, $q = 2$, and $\mu = 0$ in Lemma 1 we see that (H2)(i) implies

$$\dim N(L_2^-) = \dim N((L_2^-)^*) = 1 \quad \text{and} \quad \dim N((L_2^-)^\nu) = 2 \quad \forall \nu \geqq 2. \tag{6.4}$$

LEMMA 6. *Assume* (H0) *and* (H2)(i). *Then there exist elements* v_0, $\tilde{v}_0$, v_0^*, *and* $\tilde{v}_2^*$ *in* X_2^- *such that*

(i) $N(L_2^-) = \operatorname{span}\{v_0\}$;

(ii) $L_2^- \tilde{v}_0 = v_0$ *and* $\langle \tilde{v}_0, v_0 \rangle = 0$;

(iii) $N((L_2^*)^-) = \operatorname{span}\{v_0^*\}$ *and* $\langle v_0^*, \tilde{v}_0 \rangle = 1$;

(iv) $(L_2^*)^- \tilde{v}_0^* = 0$ *and* $\langle \tilde{v}_0^*, \tilde{v}_0 \rangle = 0$.

These elements are uniquely determined up to a scaling $(v_0, \tilde{v}_0, v_0^*, \tilde{v}_0^*) \mapsto (\beta v_0, \beta \tilde{v}_0, \beta^{-1} v_0^*, \beta^{-1} \tilde{v}_0^*)$ *with* $\beta \neq 0$; *moreover, they satisfy*

(v) $\langle v_0^*, v_0 \rangle = 0$ *and* $\langle \tilde{v}_0^*, v_0 \rangle = 1$;

(vi) $\Gamma_\sigma v_0 = \varepsilon v_0$, $\Gamma_\sigma \tilde{v}_0 = -\varepsilon \tilde{v}_0$, $\tilde{\Gamma}_\sigma v_0^* = \varepsilon v_0^*$, *and* $\tilde{\Gamma}_\sigma \tilde{v}_0^* = -\varepsilon \tilde{v}_0^*$, *with either* $\varepsilon = 1$ *or* $\varepsilon = -1$.

Proof. The proof is based on (6.4) and is completely analogous to the proof of Lemma 3; the only special point is that from $N(L_2^-) = \operatorname{span}\{v_0\} = \operatorname{span}\{\Gamma_\sigma v_0\}$ and $\Gamma_\sigma^2 v_0 = v_0$ we can only conclude that $\Gamma_\sigma v_0 = \varepsilon v_0$ with either $\varepsilon = 1$ or $\varepsilon = -1$.

Remark. Replacing $x_0(t)$ by $x_0(t + T_0/2)$ in the whole theory will change the sign of ε, so we can always assume that $\varepsilon = 1$ if we want to. However, this is not necessary for our theory, and therefore we will not use it.

We now turn to the problem of finding out how the multiplier -1 of $x_0(t)$ will change when we move along the family $\{x_\alpha\}$ of periodic solutions of (1.1) defined by (5.8). It follows from Lemma 1 that $\lambda = -e^{\mu T_\alpha}$ (with $\mu \in \mathbf{C}$) will be a Floquet multiplier of $x_\alpha(t)$ if and only if

$$L_{(1,2)}(\alpha)x = \mu x \tag{6.5}$$

has a nontrivial solution $x \in X_{(1,2)}$; in this equation $L_{(1,2)}(\alpha) \in \mathscr{L}(X_{(1,2)}, Z_{(1,2)})$ is the restriction of

$$L_2(\alpha) := D_x M_2(x^*(\alpha), \omega^*(\alpha)) \tag{6.6}$$

to $X_{(1,2)}$. To find multipliers of x_α near -1 we must solve (6.5) for μ near zero. We define $P_- \in \mathscr{L}(Z_{(1,2)})$ and $Q_- \in \mathscr{L}(Z_{(1,2)})$ by

$$P_- z := \langle \tilde{v}_0^*, z \rangle v_0, \qquad Q_- z := \langle v_0^*, z \rangle \tilde{v}_0; \tag{6.7}$$

it follows from Lemma 6 that P_- and Q_- are projections, that $R(P_-) = N(L_{(1,2)}(0))$ and $N(Q_-) = R(L_{(1,2)}(0))$, and hence that $L_{(1,2)}(0) = L_{(1,2)}$ is an isomorphism from $N(P_-)$ onto $N(Q_-)$. In (6.5) we put $x = \beta v_0 + w$, with $w \in N(P_-)$, and rewrite the equation as a system of two equations:

$$(I - Q_-)(L_{(1,2)}(\alpha) - \mu)(\beta v_0 + w) = 0, \tag{6.8}$$

$$\langle v_0^*, (L_{(1,2)}(\alpha) - \mu)(\beta v_0 + w) \rangle = 0. \tag{6.9}$$

For (α, μ) sufficiently small, (6.8) can be solved for $w = \beta w^*(\alpha, \mu)$, where $w^*(\alpha, \mu) \in N(P_-)$ is uniquely determined by

$$(I - Q_-)(L_{(1,2)}(\alpha) - \mu)(v_0 + w^*(\alpha, \mu)) = 0. \tag{6.10}$$

In particular, we have $w^*(0, 0) = 0$, while $D_\mu w^*(0, 0)$ is determined by

$$L_{(1,2)} D_\mu w^*(0, 0) = v_0 \quad \text{and} \quad D_\mu w^*(0, 0) \in N(P_-). \tag{6.11}$$

Moreover, (6.10), $L_{(1,2)}(\alpha)\Gamma_\sigma = \tilde{\Gamma}_\sigma L_{(1,2)}(\alpha)$, and $\Gamma_\sigma v_0 = \varepsilon v_0$ imply that

$$\Gamma_\sigma w^*(\alpha, \mu) = \varepsilon w^*(\alpha, -\mu). \tag{6.12}$$

Bringing the solution of (6.8) into (6.9) we see that (6.5) will have nontrivial solutions if and only if

$$G(\alpha, \mu) := \langle v_0^*, (L_{(1,2)}(\alpha) - \mu)(v_0 + w^*(\alpha, \mu)) \rangle = 0. \tag{6.13}$$

We have $G(0, 0) = 0$, while (6.12) implies that

$$G(\alpha, -\mu) = G(\alpha, \mu). \tag{6.14}$$

Also, $G(\alpha, \mu)$ is smooth in $\alpha \in \mathbf{R}$ and analytic in $\mu \in \mathbf{C}$. An easy calculation shows that

$$\begin{aligned} D_\mu^2 G(0, 0) &= -2\langle v_0^*, D_\mu w^*(0, 0)\rangle = -2\langle L_{(1,2)}^* \tilde{v}_0^*, D_\mu w^*(0, 0)\rangle \\ &= -2\langle \tilde{v}_0^*, L_{(1,2)} D_\mu w^*(0, 0)\rangle = -2\langle \tilde{v}_0^*, v_0\rangle = -2, \end{aligned} \tag{6.15}$$

while using (5.6) and (5.7) we find that

$$\begin{aligned} D_\alpha G(0, 0) &= \langle v_0^*, DL_{(1,2)}(0) \cdot v_0\rangle \\ &= \langle v_0^*, D_x^2 M_2(x_0, 1) \cdot (u_0, v_0)\rangle - \langle v_0^*, \dot{v}_0\rangle. \end{aligned} \tag{6.16}$$

More explicitly, we have

$$\begin{aligned} D_\alpha G(0, 0) = \frac{1}{2T_0} \int_0^{2T_0} &(v_0^*(t), D^2 f(x_0(t)) \cdot (u_0(t), v_0(t)))\, dt \\ &- \frac{1}{2T_0} \int_0^{2T_0} (v_0^*(t), Df(x_0(t)) \cdot v_0(t))\, dt. \end{aligned} \tag{6.17}$$

Now suppose that $D_\alpha G(0, 0) \neq 0$. Since $G(\alpha, 0)$ is real-valued it follows that $G(\alpha, 0) > 0$ if α is sufficiently small and $\alpha D_\alpha G(\alpha, 0) > 0$, whereas $G(\alpha, 0) < 0$ if $\alpha D_\alpha G(\alpha, 0) < 0$. Since $G(\alpha, \mu)$ is analytic and even in μ, it follows from (6.15) and the Weierstrass Preparation Theorem that we can write $G(\alpha, \mu)$ in the form

$$G(\alpha, \mu) = (\chi(\alpha) - \mu^2) H(\alpha, \mu), \tag{6.18}$$

with $H(\alpha, \mu)$ smooth in α, analytic and even in μ, $H(0, 0) = 1$, and with $\chi(\alpha)$ smooth and real-valued, satisfying $\chi(0) = 0$, $\chi(\alpha) > 0$ if $\alpha D_\alpha G(0, 0) > 0$ and $\chi(\alpha) < 0$ if $\alpha D_\alpha G(0, 0) < 0$. It follows that (6.13) will have two real solutions $\pm\mu_\alpha := \pm(\chi(\alpha))^{1/2}$ if $\alpha D_\alpha G(0, 0) > 0$, and two purely imaginary solutions $\pm i\tilde{\mu}_\alpha := \pm i(-\chi(\alpha))^{1/2}$ if we have $\alpha D_\alpha G(0, 0) < 0$. Via Lemma 1 this means that for one sign of α the T_α-periodic solution $x_\alpha(t)$ of (1.1) will have two real and simple multipliers near -1, while for the other sign of α we have a pair of simple multipliers on the unit circle near -1. This is the situation that should occur generically, and therefore we include the condition $D_\alpha G(0, 0) \neq 0$, which leads to this situation, in our hypotheses:

(H2) (ii) The expression $D_\alpha G(0, 0)$, as given by (6.17), is different from zero.

We can summarize our analysis up to this point as follows.

LEMMA 7. *Assume* (H0), (H1), *and* (H2). *Then the* T_α*-periodic solution* $x_\alpha(t)$ *of* (1, 1) *given by* (5.8) *has for sufficiently small* $\alpha \neq 0$ *the following multipliers near* -1:

(i) *A pair of real and simple multipliers if* $\alpha D_\alpha G(0, 0) > 0$; *these multipliers have the form* $-\exp(\pm\mu_\alpha T_\alpha)$, *with* $\mu_\alpha \sim c|\alpha|^{1/2}$ *as* $\alpha \to 0$;

(ii) *A pair of simple multipliers on the unit circle if* $\alpha D_\alpha G(0, 0) < 0$; *these multipliers have the form* $-\exp(\pm i\tilde{\mu}_\alpha T_\alpha)$, *with* $\tilde{\mu}_\alpha \sim c|\alpha|^{1/2}$ *as* $\alpha \to 0$.

Now we return to our main problem, namely, that of solving (4.12) in the case $q=2$ and under the additional hypothesis (H2). It follows from (6.1)-(6.3), (4.22), and Lemma 6 that

$$N(\tilde{L}_2)=\text{span}\,\{u_0, v_0\} \tag{6.19}$$

while

$$R(\tilde{L}_2)=\{y\in Y_2 \mid \langle u_0^*, y\rangle=\langle v_0^*, y\rangle=0\}. \tag{6.20}$$

We define in Y_2 two projection operators P and Q:

$$Py:=\langle u_0^*, y\rangle u_0+\langle \tilde{v}_0^*, y\rangle v_0 \quad \forall y\in Y_2, \tag{6.21}$$

$$Qy:=\langle u_0^*, y\rangle u_0+\langle v_0^*, y\rangle \tilde{v}_0 \quad \forall y\in Y_2. \tag{6.22}$$

The fact that P and Q are indeed projections follows from Lemmas 3 and 6 and the fact that $\langle u, v\rangle=0$ when $u\in Z_1$ and $v\in Z_2^-$. We have

$$N(\tilde{L}_2)=R(P) \quad \text{and} \quad R(\tilde{L}_2)=N(Q), \tag{6.23}$$

and we verify directly that

$$\Gamma_1 P=P\Gamma_1, \qquad \Gamma_\sigma P=P\Gamma_\sigma, \tag{6.24}$$

$$\tilde{\Gamma}_1 Q=Q\tilde{\Gamma}_1, \qquad \tilde{\Gamma}_\sigma Q=Q\tilde{\Gamma}_\sigma. \tag{6.25}$$

We use P and Q to apply a straightforward Lyapunov-Schmidt reduction to (4.21), as follows. We write $y\in X_2\cap Y_2$ as $y=\alpha u_0+\rho v_0+w$, with $\alpha\in\mathbf{R}$, $\rho\in\mathbf{R}$ and $w\in X_2\cap N(P)$. We also write (4.12) as a set of two equations:

$$(I-Q)\tilde{M}_2(\alpha u_0+\rho v_0+w)=0, \tag{6.26}$$

$$Q\tilde{M}_2(\alpha u_0+\rho v_0+w)=0. \tag{6.27}$$

In the usual way we can then solve (6.26) for $w=w^*(\alpha, \rho)$, with $w^*(0,0)=0$, $D_\alpha w^*(0,0)=D_\rho w^*(0,0)=0$, and

$$\Gamma_1 w^*(\alpha,\rho)=w^*(\alpha,-\rho), \qquad \Gamma_\sigma w^*(\alpha,\rho)=w^*(\alpha,\varepsilon\rho). \tag{6.28}$$

Bringing this solution into (6.27) gives us two scalar bifurcation equations:

$$g_0(\alpha,\rho):=\langle u_0^*, \tilde{M}_2(\alpha u_0+\rho v_0+w^*(\alpha,\rho))\rangle=0, \tag{6.29}$$

$$g(\alpha,\rho):=\langle v_0^*, \tilde{M}_2(\alpha u_0+\rho v_0+w^*(\alpha,\rho))\rangle=0. \tag{6.30}$$

The bifurcation functions g_0 and g have the following symmetry properties:

$$g_0(\alpha,-\rho)=g_0(\alpha,\rho) \quad \text{and} \quad g(\alpha,-\rho)=-g(\alpha,\rho), \tag{6.31}$$

from the Γ_1-symmetry, and

$$g_0(\alpha,\varepsilon\rho)=-g_0(\alpha,\rho), \qquad g(\alpha,\varepsilon\rho)=\varepsilon g(\alpha,\rho), \tag{6.32}$$

from the Γ_σ-symmetry.

In both cases $\varepsilon=+1$ and $\varepsilon=-1$ we see that $g_0(\alpha,\rho)\equiv 0$, i.e., (6.29) is automatically satisfied. The function $g(\alpha,\rho)$ is odd in the variable ρ, and hence we can write

$$g(\alpha,\rho)=\rho h(\alpha,\rho) \tag{6.33}$$

for some appropriate function $h(\alpha,\rho)$ satisfying

$$h(\alpha,-\rho)=h(\alpha,\rho). \tag{6.34}$$

It follows from (6.33) that setting $\rho = 0$ gives us a first solution branch of (4.12), of the form $\{\alpha u_0 + w^*(\alpha, 0) \,|\, |\alpha| < \alpha_0\}$; since $\Gamma_1 w^*(\alpha, 0) = w^*(\alpha, 0)$ this solution branch belongs in fact to $X_1 \cap Y_1$, and must therefore coincide with the branch given by Theorem 2, i.e., we have

$$x_0 + \alpha u_0 + w^*(\alpha, 0) = x^*(\alpha). \tag{6.35}$$

To find other solutions we must set $\rho \neq 0$, and then (6.30) reduces via (6.33) to

$$h(\alpha, \rho) = 0. \tag{6.36}$$

We have (using (6.35))

$$\begin{aligned} h(\alpha, 0) = D_\rho g(\alpha, 0) &= \langle v_0^*, D\tilde{M}_2(\alpha u_0 + w^*(\alpha, 0)) \cdot (v_0 + D_\rho w^*(\alpha, 0))\rangle \\ &= \langle v_0^*, D_x M_2(x^*(\alpha), \omega^*(\alpha)) \cdot (v_0 + D_\rho w^*(\alpha, 0))\rangle; \end{aligned} \tag{6.37}$$

it follows that

$$h(0, 0) = \langle v_0^*, L_2 v_0\rangle = 0 \tag{6.38}$$

and

$$D_\alpha h(0, 0) = \langle v_0^*, D_\alpha L_2(0) v_0\rangle. \tag{6.39}$$

Comparing with (6.16) we see that $D_\alpha h(0, 0) \neq 0$ by (H2)(ii). Therefore we can solve (6.36) for $\alpha = \alpha^*(\rho)$, with $\alpha^*(0) = 0$ and $\alpha^*(-\rho) = \alpha^*(\rho)$. This gives us a solution branch

$$\{\tilde{y}(\rho) := \alpha^*(\rho) u_0 + \rho v_0 + w^*(\alpha^*(\rho), \rho) \,|\, |\rho| < \rho_0\}$$

for (4.12). Since $\tilde{y}(-\rho) = \Gamma_1 \tilde{y}(\rho)$ both solutions $\tilde{y}(\rho)$ and $\tilde{y}(-\rho)$ lead to the same group orbit of solutions for (3.5). Moreover, we have

$$\Gamma_\sigma \tilde{y}(\rho) = \tilde{y}(\varepsilon\rho), \tag{6.40}$$

i.e., $\Gamma_\sigma \tilde{y}(\rho) = \tilde{y}(\rho)$ if $\varepsilon = 1$, and $\Gamma_\sigma \Gamma_1 \tilde{y}(\rho) = \tilde{y}(\rho)$ if $\varepsilon = -1$. It follows that the solutions $\tilde{y}(\rho)$ are symmetric in the sense defined in § 1. Putting

$$x_d^*(\rho) := x_0 + \tilde{y}(\rho) \quad \text{and} \quad \omega_d^*(\rho) := \tilde{\omega}(\tilde{y}(\rho)) \tag{6.41}$$

we therefore have the following result.

THEOREM 3. *Assume* (H0)-(H2). *Then there exist a neighborhood* $\mathcal{U}$ *of* $\kappa \times \{1\}$ *in* $X_2 \times \mathbf{R}$, *numbers* $\alpha_0 > 0$ *and* $\rho_0 > 0$, *and smooth mappings* $x_d^*:]-\rho_0, \rho_0[\to X_2$ *and* $\omega_d^*:]-\rho_0, \rho_0[\to \mathbf{R}$, *such that the following holds*:

(i) $\{(x, \omega) \in \mathcal{U} \,|\, M_2(x, \omega) = 0\} = \{(\Gamma_\phi \cdot x^*(\alpha), \omega^*(\alpha)) \,|\, |\alpha| < \alpha_0, \phi \in S^1\}$
$\cup \{(\Gamma_\phi \cdot x_d^*(\rho), \omega_d^*(\rho)) \,|\, 0 < \rho < \rho_0, \phi \in S^1\}$,

where $x^*(\alpha)$ *and* $\omega^*(\alpha)$ *are as in Theorem* 2;

(ii) $(x_d^*(0), \omega_d^*(0)) = (x_0, 1)$;

(iii) $\langle \tilde{v}_0^*, x_d^*(\rho) - x_0\rangle = \rho$, $\langle u_0^*, x_d^*(\rho) - x_0\rangle = 0$ *and* $\langle \tilde{u}_0^*, x_d^*(\rho) - x_0\rangle = 0$;

(iv) *For* $\rho \neq 0$, $x_d^*(\rho)$ *has minimal period* $2T_0$;

(v) *Either* $\Gamma_\sigma x_d^*(\rho) = x_d^*(\rho)$ *for all* ρ, *or* $\Gamma_\sigma \Gamma_1 x_d^*(\rho) = x_d^*(\rho)$ *for all* ρ;

(vi) $\omega_d^*(-\rho) = \omega_d^*(\rho)$.

Setting

$$x_\rho^d(t) := x_d^*(\rho) \cdot (\omega_d^*(\rho) t) \tag{6.42}$$

it follows from Theorem 3 that $\{x_\rho^d(t) \,|\, 0 < \rho < \rho_0\}$ forms a branch of symmetric periodic solutions of (1.1), converging to $x_0(t)$ as $\rho \to 0$, and with minimal period

$$T_\rho^d := \frac{2T_0}{\omega_d^*(\rho)} \tag{6.43}$$

converging to $2T_0$ as $\rho \to 0$. So we have a period-doubling branching at x_0. When $D^2\omega_d^*(0) \neq 0$, as will generically be the case, then

$$\frac{dT_\rho^d}{d\rho} \neq 0 \quad \text{for } \rho > 0, \tag{6.44}$$

and, by the arguments of § 2, the T_ρ^d-periodic solution $x_\rho^d(t)$ of (1.1) will have 1 as a non-semisimple Floquet multiplier with algebraic multiplicity 2. It is possible to calculate $D^2\omega_d^*(0)$ from our foregoing formalism, but this is a lengthy and tedious exercise that we do not want to include here.

7. The case $q \geqq 3$. In this section we fix some integer $q \geqq 3$, and together with (H0) and (H1) we also assume:

(Hq) (i) $x_0(t)$ has a pair of simple characteristic multipliers $(\lambda_q^p, \bar{\lambda}_q^p)$, for some integer p with $0 < p < q$ and greatest common divisor (g.c.d.) $(p, q) = 1$;
(ii) (nonresonance) $x_0(t)$ has, besides 1, λ_q^p, and $\bar{\lambda}_q^p$, no other multipliers λ such that $\lambda^q = 1$.

As in § 6, we will later complement (Hq) with an appropriate transversality condition (Hq)(iii). It follows from g.c.d. $(p, q) = 1$ that there exists some integer r such that

$$rp = 1 \pmod{q}. \tag{7.1}$$

Using Lemma 1 we see that (Hq)(i) implies that

$$\dim N((L_{(p,q)})^\nu) = \dim N((L_{(p,q)}^*)^\nu) = 1 \quad \forall \nu \geqq 1. \tag{7.2}$$

LEMMA 8. *Assume* (H0) *and* (Hq)(i). *Then there exist elements* $\zeta_0 \in X_{(p,q)}$ *and* $\zeta_0^* \in X_{(p,q)}$ *such that*

(i) $N(L_{(p,q)}) = \text{span}\{\zeta_0\}$ *and* $N(L_{(p,q)}^*) = \text{span}\{\zeta_0^*\}$;
(ii) $\langle \zeta_0^*, \zeta_0 \rangle = 2$ *and* $\langle \zeta_0^*, \bar{\zeta}_0 \rangle = 0$;
(iii) $\Gamma_\sigma \zeta_0 = \bar{\zeta}_0$ *and* $\tilde{\Gamma}_\sigma \zeta_0^* = -\bar{\zeta}_0^*$.

Proof. From (7.2) there exist elements ζ and ζ^* of $X_{(p,q)}$ such that $N(L_{(p,q)}) = \text{span}\{\zeta\}$ and $N(L_{(p,q)}^*) = \text{span}\{\zeta^*\}$. But also $\Gamma_\sigma \bar{\zeta}$ belongs to $X_{(p,q)}$, and from $L_q \Gamma_\sigma = \tilde{\Gamma}_\sigma L_q$ and the fact that L_q is a real operator it follows that $\Gamma_\sigma \bar{\zeta} \in N(L_{(p,q)})$. So we have $\Gamma_\sigma \bar{\zeta} = \beta\zeta$ for some $\beta \in \mathbf{C}$, which then must satisfy $|\beta| = 1$, i.e., $\beta = \exp(2i\phi)$ for some $\phi \in \mathbf{R}$. Setting $\zeta_0 = \zeta \exp(i\phi)$ it follows that $N(L_{(p,q)}) = \text{span}\{\zeta_0\}$ and $\Gamma_\sigma \zeta_0 = \bar{\zeta}_0$. It follows from (7.2) that $\zeta_0 \notin R(L_{(p,q)})$, and therefore we have $\langle \zeta^*, \zeta_0 \rangle \neq 0$, by (3.35). Taking $\zeta_0^* := 2\overline{\langle \zeta^*, \zeta_0 \rangle}^{-1} \zeta^*$ it follows that $N(L_{(p,q)}^*) = \text{span}\{\zeta_0^*\}$ and $\langle \zeta_0^*, \zeta_0 \rangle = 2$. The same argument as above shows that $\tilde{\Gamma}_\sigma \bar{\zeta}_0^* = \beta\zeta_0^*$ for some $\beta \in \mathbf{C}$; but then we have

$$2 = \langle \zeta_0^*, \Gamma_\sigma \bar{\zeta}_0 \rangle = -\langle \tilde{\Gamma}_\sigma \zeta_0^*, \bar{\zeta}_0 \rangle = -\beta \langle \bar{\zeta}_0^*, \bar{\zeta}_0 \rangle = -2\beta,$$

and hence $\beta = -1$. Finally, we remark that $\zeta_0 \in Z_{(p,q)}$ implies that $\bar{\zeta}_0 \in Z_{(-p,q)} = Z_{(q-p,q)}$, and hence $\langle \zeta_0^*, \bar{\zeta}_0 \rangle = 0$ by (3.32) and $q - p \neq p \pmod{q}$ (this follows from g.c.d. $(p, q) = 1$).

Next we look for the multipliers of the T_α-periodic solution $x_\alpha(t)$ of (1.1) given by (5.8). By Lemma 1 we see that $\lambda := \lambda_p^q e^{i\mu T_\alpha}$ will be a multiplier of x_α is and only if

$$L_{(p,q)}(\alpha)x = i\mu x \tag{7.3}$$

has a nontrivial solution $x \in X_{(p,q)}$; in (7.3) $L_{(p,q)}(\alpha) \in \mathscr{L}(X_{(p,q)}, Z_{(p,q)})$ is the restriction to $X_{(p,q)}$ of the operator $L_q(\alpha)^c$, where $L_q(\alpha) \in \mathscr{L}(X_q, Z_q)$ is defined by

$$L_q(\alpha) := D_x M_q(x^*(\alpha), \omega^*(\alpha)). \tag{7.4}$$

So to find multipliers of $x_\alpha(t)$ near λ_q^p we must solve (7.3) for μ near zero. To do so we define an operator $\pi \in \mathscr{L}(Z_{(p,q)})$ by

$$\pi z := \tfrac{1}{2}\langle \zeta_0^*, z\rangle \zeta_0 \quad \forall z \in Z_{(p,q)}; \tag{7.5}$$

It follows from Lemma 8 that π is a projection, with

$$R(\pi) = N(L_{(p,q)}(0)) \quad \text{and} \quad N(\pi) = R(L_{(p,q)}(0)). \tag{7.6}$$

We put $x = \beta\zeta_0 + w$ in (7.3), with $\beta \in \mathbf{C}$ and $w \in N(\pi) \cap X_{(p,q)}$, and rewrite the equation as a set of two equations:

$$(I-\pi)(L_{(p,q)}(\alpha) - i\mu)(\beta\zeta_0 + w) = 0, \tag{7.7}$$

$$\langle \zeta_0^*, (L_{(p,q)}(\alpha) - i\mu)(\beta\zeta_0 + w)\rangle = 0. \tag{7.8}$$

For (α, μ) sufficiently small, (7.7) can be solved for $w = \beta w^*(\alpha, \mu)$, where $w^*(\alpha, \mu) \in N(\pi)$ is uniquely determined by

$$(I-\pi)(L_{(p,q)}(\alpha) - i\mu)(\zeta_0 + w^*(\alpha, \mu)) = 0. \tag{7.9}$$

In particular, $w^*(\alpha, \mu)$ is smooth in α and analytic in μ, with $w^*(0,0) = 0$ and $D_\mu w^*(0,0) = 0$. We can also verify from (7.5) that $\pi\tilde{\Gamma}_\sigma \bar{z} = \tilde{\Gamma}_\sigma(\overline{\pi z})$; then the uniqueness of the solution of (7.9) implies that

$$\Gamma_\sigma \overline{w^*(\alpha, \mu)} = w^*(\alpha, \bar{\mu}). \tag{7.10}$$

Bringing the solution of (7.7) into (7.8) we see that (7.3) will have nontrivial solutions if and only if

$$H(\alpha, \mu) := \langle \zeta_0^*, (L_{(p,q)}(\alpha) - i\mu)(\zeta_0 + w^*(\alpha, \mu))\rangle = 0. \tag{7.11}$$

Again, $H(\alpha, \mu)$ is smooth in α and analytic in μ, with $H(0,0) = 0$ and $D_\mu H(0,0) = -i\langle \zeta_0^*, \zeta_0\rangle = -2i$. Therefore we can solve (7.11) for $\mu = \mu^*(\alpha)$, with $\mu^*(0) = 0$. Using (7.10) and (7.11), we also have that

$$H(\alpha, \bar{\mu}) = -\overline{H(\alpha, \mu)};$$

the uniqueness of the solution of (7.11) then implies that

$$\overline{\mu^*(\alpha)} = \mu^*(\alpha), \tag{7.12}$$

i.e., $\mu^*(\alpha) \in \mathbf{R}$. Finally, we have that $D\mu^*(0) \neq 0$ if and only if $D_\alpha H(0,0) \neq 0$. It follows from (7.11) that

$$D_\alpha H(0,0) = \langle \zeta_0^*, DL_{(p,q)}(0) \cdot \zeta_0\rangle = \langle \zeta_0^*, DL_q(0) \cdot \zeta_0\rangle, \tag{7.13}$$

or more explicitly, using the results of § 5:

$$\begin{aligned} D_\alpha H(0,0) &= \langle \zeta_0^*, D_x^2 M_q(x_0, 1) \cdot (u_0, \zeta_0)\rangle - \langle \zeta_0^*, \dot{\zeta}_0\rangle \\ &= \frac{1}{qT_0}\int_0^{qT_0} (\zeta_0^*(t), D^2 f(x_0(t))) \cdot (u_0(t), \zeta_0(t))\, dt \\ &\quad - \frac{1}{qT_0}\int_0^{qT_0} (\zeta_0^*(t), Df(x_0(t)) \cdot \zeta_0(t))\, dt. \end{aligned} \tag{7.14}$$

LEMMA 9. *Assume* (H0), (H1), *and* (Hq)(i). *Then the T_α-periodic solution $x_\alpha(t)$ of* (1.1) *has, for α sufficiently small, a unique pair of simple multipliers on the unit circle*

and near $(\lambda_q^p, \bar{\lambda}_q^p)$. *These multipliers have the form* $\exp(\pm i\psi^*(\alpha))$, *with* $\alpha \mapsto \psi^*(\alpha)$ *a smooth real-valued mapping,* $\psi^*(0) = 2\pi p/q$, *and* $D\psi^*(0) \neq 0$ *if and only if* $D_\alpha H(0,0) \neq 0$.

Proof. Take $\psi^*(\alpha) = 2\pi p/q + \mu^*(\alpha) T_\alpha$.

So, if $D_\alpha H(0,0) \neq 0$, then the multipliers $\exp(\pm i\psi^*(\alpha))$ of $x_\alpha(t)$ move with nonzero speed along the unit circle; since this is the situation which should happen generically we make it an explicit hypothesis:

(Hq) (iii) The expression $D_\alpha H(0,0)$, as given by (7.14), is different from zero.

Now we return to our main problem, which is to solve (4.12) for y near zero, under the hypothesis (Hq). It follows from (4.20)-(4.22), (3.35), and Lemma 8 that

$$N(\tilde{L}_q) = \text{span}\{u_0, \text{Re } \zeta_0, \text{Im } \zeta_0\}, \tag{7.15}$$

$$R(\tilde{L}_q) = \{y \in Y_q \mid \langle u_0^*, y\rangle = 0 \text{ and } \langle \zeta_0^*, y\rangle = 0\}. \tag{7.16}$$

We define a projection $P \in \mathscr{L}(Y_q)$ by

$$Py := \langle u_0^*, y\rangle u_0 + \text{Re}\,(\langle \zeta_0^*, y\rangle \zeta_0) \quad \forall y \in Y_q. \tag{7.17}$$

Then we have

$$R(P) = N(\tilde{L}_q) \quad \text{and} \quad N(P) = R(\tilde{L}_q), \tag{7.18}$$

while also

$$P\Gamma_1 = \Gamma_1 P \quad \text{and} \quad P\Gamma_\sigma = \Gamma_\sigma P. \tag{7.19}$$

We use P to perform a Lyapunov-Schmidt reduction on (4.12). We write $y \in Y_q \cap X_q$ in the form

$$y = \alpha u_0 + \text{Re}\,(z\zeta_0) + v,$$

with $\alpha \in \mathbf{R}$, $z \in \mathbf{C}$, and $v \in N(P) \cap X_q$. We should remark here that although we use complex coordinates ($z \in \mathbf{C}$), this is only for notational convenience, and we should consider $\mathbf{C}$ as a two-dimensional real vector space; this is particularly true when we talk about smoothness (see further). We can now rewrite (4.12) as a system of two equations:

$$(I - P)\tilde{M}_q(\alpha u_0 + \text{Re}\,(z\zeta_0) + v) = 0, \tag{7.20}$$

$$P\tilde{M}_q(\alpha u_0 + \text{Re}\,(z\zeta_0) + v) = 0. \tag{7.21}$$

Equation (7.20) can be solved for $v = v^*(\alpha, z)$, with $v^*(0,0) = 0$. The mapping $v^*: \mathbf{R} \times \mathbf{C} \to N(P) \cap X_q$ is smooth near the origin (in the sense explained above), and we also have that

$$\Gamma_1 v^*(\alpha, z) = v^*(\alpha, \lambda_q^p z), \tag{7.22}$$

$$\Gamma_\sigma v^*(\alpha, z) = v^*(\alpha, \bar{z}). \tag{7.23}$$

Bringing the solution of (7.20) into (7.21) and using (7.17) gives us two bifurcation equations:

$$g_0(\alpha, z) := \langle u_0^*, \tilde{M}_q(\alpha u_0 + \text{Re}\,(z, \zeta_0) + v^*(\alpha, z))\rangle = 0 \tag{7.24}$$

$$g(\alpha, z) := \langle \zeta_0^*, \tilde{M}_q(\alpha u_0 + \text{Re}\,(z\zeta_0) + v^*(\alpha, z))\rangle = 0. \tag{7.25}$$

The mappings $g_0: \mathbf{R} \times \mathbf{C} \to \mathbf{R}$ and $g: \mathbf{R} \times \mathbf{C} \to \mathbf{C}$ are smooth near the origin, with $g_0(0,0) = 0$ and $g(0,0) = 0$. They also have the following symmetry properties:

$$g_0(\alpha, \lambda_q^p z) = g_0(\alpha, z), \qquad g(\alpha, \lambda_q^p z) = \lambda_q^p g(\alpha, z), \tag{7.26}$$

from the Γ_1-symmetry, and

$$g_0(\alpha, \bar{z}) = -g_0(\alpha, z), \qquad g(\alpha, \bar{z}) = -\overline{g(\alpha, z)}, \tag{7.27}$$

from the Γ_σ-symmetry.

LEMMA 10. *There exist smooth functions* $h_i : \mathbf{R} \times \mathbf{C} \to \mathbf{R}$ $(i = 0, 1, 2)$ *such that*

(i) $g_0(\alpha, z) = h_0(\alpha, z)$ Im $\bar{z}^q$;

(ii) $g(\alpha, z) = ih_1(\alpha, z)z + ih_2(\alpha, z)\bar{z}^{q-1}$;

(iii) $h_i(\alpha, \lambda_q z) = h_i(\alpha, \bar{z}) = h_i(\alpha, z)$, $(i = 0, 1, 2)$.

The proof of this lemma is completely analogous to the proof of the Lemma 6 in the paper [8], and therefore we do not repeat it here. For a different approach, see Golubitsky, Stewart, and Schaeffer [3]. The results in [3] and [8] also show that (iii) implies that the functions $h_i(\alpha, z)$ must have the form

$$h_i(\alpha, z) = \sum_{l=0}^{q'} A_{i,l}(\alpha)|z|^{2l} + O(|z|^q) \quad \text{as } z \to 0, \tag{7.28}$$

where $q' := [(q-1)/2]$ and the functions $A_{i,l} : \mathbf{R} \to \mathbf{R}$ are smooth $(i = 0, 1, 2;\ l = 0, 1, \cdots, q')$.

It follows immediately from Lemma 10 that (7.24), (7.25) are satisfied for $z = 0$. Since the corresponding solutions of (3.5) belong to X_1, they must coincide with the solutions given by Theorem 2, i.e., we have

$$x_0 + \alpha u_0 + v^*(\alpha, 0) = x^*(\alpha). \tag{7.29}$$

Replacing $g_0(\alpha, z)$ and $g(\alpha, z)$ by their expressions from Lemma 10, we see that we can also satisfy (7.24) by taking $z \in \mathbf{C}$ such that

$$\text{Im } \bar{z}^q = 0. \tag{7.30}$$

More strongly, if $h_0(0, 0) \neq 0$ or $h_2(0, 0) \neq 0$ then all sufficiently small solutions (α, z) of (7.24), (7.25) must satisfy (7.30): this follows from the form of $g_0(\alpha, z)$, and from the fact that if we multiply (7.25) by $\bar{z}$ and take the real part of the resulting equation, then we find

$$h_2(\alpha, z) \text{ Im } \bar{z}^q = 0. \tag{7.31}$$

In principle it is possible to calculate $h_0(0, 0)$ and $h_2(0, 0)$; the result will involve higher-order terms (up to order q) of $f(x)$ at the basic solution $x_0(t)$. Generically these expressions will be nonzero, and therefore all sufficiently small solutions of (7.24), (7.25) will, in a generic situation, satisfy (7.30).

Now (7.30) means that z must have the form

$$z = \rho\lambda_q^i \quad \text{or} \quad z = \rho\lambda_{2q}^{2j+1}, \quad \rho \in \mathbf{R}, \quad j = 0, 1, \cdots, q-1. \tag{7.32}$$

Using (7.1) and (7.22) we see that

$$\alpha u_0 + \text{Re } (\lambda_q z\zeta_0) + v^*(\alpha, \lambda_q z) = \Gamma_r(\alpha u_0 + \text{Re } (z\zeta_0) + v^*(\alpha, z)),$$

and therefore different values of j in (7.32) will correspond to solutions of (3.5) on the same group orbit. It is therefore sufficient to consider the case $z = \rho \in \mathbf{R}$ if q is odd, and the cases $z = \rho$ and $z = \rho\lambda_{2q}$ with $\rho > 0$ if q is even. Moreover, we have for $\rho \in \mathbf{R}$ that

$$\alpha u_0 + \text{Re } (\rho\zeta_0) + v^*(\alpha, \rho) = \Gamma_\sigma(\alpha u_0 + \text{Re } (\rho\zeta_0) + v^*(\alpha, \rho)),$$

$$\alpha u_0 + \text{Re } (\rho\lambda_{2q}\zeta_0) + v^*(\alpha, \rho\lambda_{2q}) = \Gamma_r\Gamma_\sigma(\alpha u_0 + \text{Re } (\rho\lambda_{2q}\zeta_0) + v^*(\alpha, \rho\lambda_{2q})).$$

We conclude that the corresponding solutions of (1.1) will all be symmetric. For $z=\rho\in\mathbf{R}$, (7.24), (7.25) reduce to just one real and scalar equation, namely,

$$h_1(\alpha,\rho)+h_2(\alpha,\rho)\rho^{q-2}=0, \tag{7.33}$$

whereas for $z=\rho\lambda_{2q}$ ($\rho\in\mathbf{R}$) we find the equation

$$h_1(\alpha,\rho\lambda_{2q})-h_2(\alpha,\rho\lambda_{2q})\rho^{q-2}=0. \tag{7.34}$$

To see how we can solve these we must calculate $h_1(0,0)$ and $D_\alpha h_1(0,0)$. Combining Lemma 10 with (7.25) and taking $z=\rho\in\mathbf{R}$, we find that

$$ih_1(\alpha,\rho)\rho+ih_2(\alpha,\rho)\rho^{q-1}=\langle\zeta_0^*,\tilde{M}_q(\alpha u_0+\rho\ \mathrm{Re}\ \zeta_0+v^*(\alpha,\rho))\rangle. \tag{7.35}$$

Differentiating in ρ at $\rho=0$ and using (7.29) gives us

$$\begin{aligned} ih_1(\alpha,0)&=\langle\zeta_0^*,D\tilde{M}_q(\alpha u_0+v^*(\alpha,0))\cdot(\mathrm{Re}\ \zeta_0+D_\rho v^*(\alpha,0))\rangle\\ &=\tfrac{1}{2}\langle\zeta_0^*,L_q(\alpha)\cdot(\zeta_0+2D_\rho v^*(\alpha,0))\rangle. \end{aligned} \tag{7.36}$$

Now $v^*(0,0)=0$, $D_\alpha v^*(0,0)=0$ and $D_\rho v^*(0,0)=0$. Therefore it follows from (7.36) that $h_1(0,0)=0$, while

$$2iD_\alpha h_1(0,0)=\langle\zeta_0^*,DL_q(0)\cdot\zeta_0\rangle. \tag{7.37}$$

Comparing with (7.13) we see that $D_\alpha h_1(0,0)\neq 0$ by our hypothesis (Hq)(iii). So we can solve both equations (7.33) and (7.34) for α as a function of ρ.

Equation (7.33) gives us a solution branch $\alpha=\alpha_q(\rho)$, with $\alpha_q(0)=0$ and

$$\alpha_q(-\rho)=\alpha_q(\rho)+O(\rho^{q-2})\quad\text{as }\rho\to 0. \tag{7.38}$$

This last relation follows from the form (7.28) of the functions $h_i(\alpha,z)$. In fact, if q is even then Lemma 10(iii) implies that $h_i(\alpha,-\rho)=h_i(\alpha,\rho)$ for $\rho\in\mathbf{R}$, and then we have

$$\alpha_q(-\rho)=\alpha_q(\rho). \tag{7.39}$$

In the case where q is even we also must solve (7.34), which gives us a solution branch $\alpha=\tilde{\alpha}_q(\rho)$, with $\tilde{\alpha}_q(0)=0$ and

$$\tilde{\alpha}_q(\rho)=\alpha_q(\rho)+O(\rho^{q-2})\quad\text{as }\rho\to 0. \tag{7.40}$$

It follows from (7.38) and (7.40) that

$$D\alpha_q(0)=D\tilde{\alpha}_q(0)=0 \tag{7.41}$$

if $q\geqq 4$. In the case $q=3$ a direct calculation gives

$$D\alpha_3(0)=-\tfrac{1}{4}(D_\alpha H(0,0))^{-1}\langle\zeta_0^*,D_x^2M_3(x_0,1)\cdot(\bar{\zeta}_0,\bar{\zeta}_0)\rangle. \tag{7.42}$$

Setting

$$\begin{aligned} x_q^*(\rho)&:=x_0+\alpha_q(\rho)u_0+\rho\ \mathrm{Re}\ \zeta_0+v^*(\alpha_q(\rho),\rho),\\ \tilde{x}_q^*(\rho)&:=x_0+\tilde{\alpha}_q(\rho)u_0+\rho\ \mathrm{Re}\ (\lambda_{2q}\zeta_0)+v^*(\tilde{\alpha}_q(\rho),\rho\lambda_{2q}),\\ \omega_q^*(\rho)&:=\tilde{\omega}(x_q^*(\rho)-x_0),\\ \tilde{\omega}^*(\rho)&:=\tilde{\omega}(\tilde{x}_q^*(\rho)-x_0), \end{aligned}$$

we can summarize our results in the following theorems.

THEOREM 4. *Assume* (H0), (H1), *and* (Hq) *for some odd* $q \geqq 3$. *Then there exist a neighborhood* $\mathcal{U}$ *of* $\kappa \times \{1\}$ *in* $X_q \times \mathbf{R}$, *numbers* $\alpha_0 > 0$ *and* $\rho_0 > 0$, *and smooth mappings* $x_q^*;]-\rho_0, \rho_0[\to X_q$ *and* $\omega_q^* :]-\rho_0, \rho_0[\to \mathbf{R}$ *such that the following holds*:

$$\text{(i)}\quad \{(x, \omega) \in \mathcal{U} \mid M_q(x, \omega) = 0\} \supset \{(\Gamma_\phi \cdot x^*(\alpha), \omega^*(\alpha)) \mid |\alpha| < \alpha_0, \phi \in S^1\}$$
$$\cup \{(\Gamma_\phi \cdot x_q^*(\rho), \omega_q^*(\rho)) \mid |\rho| < \rho_0, \phi \in S^1\},$$

where $x^*(\alpha)$ *and* $\omega^*(\alpha)$ *are as in Theorem* 2, *and where generically we can replace the inclusion by an equality.*

(ii) $(x_q^*(0), \omega_q^*(0)) = (x_0, 1)$.

(iii) $Dx_q^*(0) = \operatorname{Re} \zeta_0$ *and* $D\omega_q^*(0) = 0$, *except if* $q = 3$ *in which case we have*

$$Dx_3^*(0) = \operatorname{Re} \zeta_0 + cu_0 \quad \text{and} \quad D\omega_3^*(0) = c,$$

with $c = -\frac{1}{4}(D_\alpha H(0, 0))^{-1} \langle \zeta_0^*, D_x^2 M_3(x_0, 1) \cdot (\bar{\zeta}_0, \bar{\zeta}_0) \rangle$.

(iv) $x_q^*(\rho)$ *has minimal period* qT_0 *for* $\rho \neq 0$.

(v) $T_\sigma x_q^*(\rho) = x_q^*(\rho)$.

THEOREM 5. *Assume* (H0), (H1), *and* (Hq) *for some even* $q > 3$. *Then there exist a neighborhood* $\mathcal{U}$ *of* $\kappa \times \{1\}$ *in* $X_q \times \mathbf{R}$, *numbers* $\alpha_0 > 0$ *and* $\rho_0 > 0$, *and smooth mappings* $x_q^*;]-\rho_0, \rho_0[\to X_q$, $\omega_q^* :]-\rho_0, \rho_0[\to \mathbf{R}$, $\tilde{x}_q^* :]-\rho_0, \rho_0[\to X_q$ *and* $\tilde{\omega}_q^* :]-\rho_0, \rho_0[\to \mathbf{R}$ *such that the following holds*:

$$\text{(i)}\quad \{(x, \omega) \in \mathcal{U} \mid M_q(x, \omega) = 0\} \supset \{(\Gamma_\phi \cdot x^*(\alpha), \omega^*(\alpha)) \mid |\alpha| < \alpha_0, \phi \in S^1\}$$
$$\cup \{(\Gamma_\phi \cdot x_q^*(\rho), \omega_q^*(\rho)) \mid 0 < \rho < \rho_0, \phi \in S^1\}$$
$$\cup \{(\Gamma_\phi \cdot \tilde{x}_q^*(\rho), \tilde{\omega}_q^*(\rho)) \mid 0 < \rho < \rho_0, \phi \in S^1\},$$

where $x^*(\alpha)$ *and* $\omega^*(\alpha)$ *are as in Theorem* 2, *and where generically we can replace the inclusion by an equality.*

(ii) $(x_q^*(0), \omega_q^*(0)) = (\tilde{x}_q^*(0), \tilde{\omega}_q^*(0)) = (x_0, 1)$.

(iii) $Dx_q^*(0) = \operatorname{Re} \zeta_0$, $D\tilde{x}_q^*(0) = \operatorname{Re}(\lambda_{2q}\zeta_0)$, $D\omega_q^*(0) = 0$, *and* $D\tilde{\omega}_q^*(0) = 0$.

(iv) $x^*(q)$ *and* $\tilde{x}_q^*(\rho)$ *have minimal period* qT_0 *for* $\rho \neq 0$.

(v) $\Gamma_\sigma x_q^*(\rho) = x_q^*(\rho)$ *and* $\Gamma_r \Gamma_\sigma \tilde{x}_q^*(\rho) = \tilde{x}_q^*(\rho)$.

We conclude from Theorems 4 and 5 that under (Hq) we have subharmonic branching at the solution $x_0(t)$ of (1.1). If q is odd, then two branches of symmetric periodic solutions pass through x_0: a primary branch, along which the minimal period converges to T_0 as we approach x_0, and a secondary branch along which the minimal period approaches qT_0 as we approach x_0. In the case where q is even, next to the primary branch there are two other branches of symmetric periodic solutions that terminate at x_0; along each of these secondary branches the period converges to qT_0 as we approach x_0.

Acknowledgment. I thank Professor G. Iooss for his hospitality and for a number of interesting discussions concerning this paper.

REFERENCES

[1] E. BREDON, *Introduction to Compact Transformation Groups*, Academic Press, New York, 1972.

[2] R. DEVANEY, *Reversible diffeomorphisms and flows*, Trans. Amer. Math. Soc., 218 (1976), pp. 89-113.

[3] M. GOLUBITSKY, I. STEWART, AND D. SCHAEFFER, *Singularities and Groups in Bifurcation Theory, Vol.* II, Appl. Math. Sci. 69, Springer-Verlag, New York, 1988.

[4] K. MEYER, *Generic bifurcation of periodic points*, Trans. Amer. Math. Soc., 149 (1970), pp. 95-107.

[5] M. SEVRYUK, *Reversible Systems*, Lecture Notes in Math. 1211, Springer-Verlag, New York, 1986.

[6] A. VANDERBAUWHEDE, *Secondary bifurcations of periodic solutions in autonomous systems*, C.M.S. Conf. Proc. 8, American Mathematical Society, Providence, RI, 1987, pp. 693–701.

[7] ———, *Period-doublings and orbit-bifurcations in symmetric systems*, in Proc. Banach Institute, Warsaw, to appear.

[8] ———, *Bifurcation of subharmonic solutions in time-reversible systems*, J. Appl. Math. Phys., 37 (1986), pp. 455–478.

[9] ———, *Local Bifurcation and Symmetry*, Research Notes in Math. 75, Pitman, London, 1982.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 980-994, July 1990
© 1990 Society for Industrial and Applied Mathematics
009

ON THE STURM-PICONE THEOREM FOR nth-ORDER DIFFERENTIAL EQUATIONS*

MARCELLINO GAUDENZI†

Abstract. The classical Sturm-Picone theorem can be seen as a comparison theorem for the two-point boundary value problem (BVP) associated with the linear equation $(py')'+qy=0$. In this paper the problem of extending this theorem to two-point BVPs associated with the equation $p_n(p_{n-1}\cdots(p_1(p_0y)')'\cdots)'+qy=0$ is considered. It is shown that the extension is possible only for particular types of boundary conditions. In the case $n=3$, the BVPs for which the extension is true are characterized. Analogous results are obtained for the corresponding eigenvalue problems.

Key words. linear differential equations of arbitrary order, boundary value problems, eigenvalue comparison

AMS(MOS) subject classification. primary 34C10

Introduction. During a visit at the International School for Advanced Studies (SISSA) in Trieste, Prof. A. C. Lazer raised the following question: Let us consider the pair of linear differential equations

$$(p(x)y^{(m)})^{(3-m)}+q(x)y=0, \tag{1}$$

$$(\tilde{p}(x)y^{(m)})^{(3-m)}+q(x)y=0, \tag{2}$$

with associated boundary conditions

$$y(a)=y'(a)=0, \qquad y(b)=0, \tag{1.1}$$

$$y(a)=y'(a)=0, \qquad y(c)=0, \tag{2.1}$$

where $m=1$ or $m=2$, p and $\tilde{p}$ are positive functions of class C^{3-m} on the interval $[a, b]$, and q is a nonnegative continuous function defined on $[a, b]$. Is it true that if $p(x)\geqq\tilde{p}(x)>0$ on $[a, b]$ and $p\not\equiv\tilde{p}$, then the existence of a nontrivial solution of the boundary value problem (BVP) (1), (1.1) implies the existence of a nontrivial solution of the BVP (2), (2.1) for some $c<b$?

An affirmative answer to this question would extend the Sturm-Picone comparison theorem to equations of higher order. Such problems have already been studied in the case of the comparison with respect to changing the function q (see [1], [2], [11]).

In this paper we prove some general results about the validity or nonvalidity of the Sturm-Picone comparison theorem for nth-order equations. Analogous results on the comparison of the first positive eigenvalue are also obtained.

In particular, as consequence of our main theorems, we have:

(1) A negative answer to the problem raised by Lazer in both cases $m=1$ and $m=2$;

(2) A positive answer to Lazer's question if we change the boundary conditions properly (e.g., $m=1$ and $y(a)=y'(a)=0$, $y'(b)=0$);

(3) A characterization of the third-order BVPs for which the extension of the Sturm-Picone theorem holds.

* Received by the editors May 1, 1987; accepted for publication (in revised form) August 1, 1989.
† Dipartimento di Matematica ed Informatica, via Zanon 6, 33100 Udine, Italy.

Moreover, as a byproduct of our negative results we provide counterexamples showing that Theorem 1.3 of [9] is not correct.

1. Preliminaries. In the sequel we will consider an equation that contains as a special case (1), (2), and the Sturm-Picone equation. We summarize here the notation used and well-known facts about this equation that we will use in the next sections.

Let us give the linear nth-order equation

$$p_n(p_{n-1}\cdots(p_1(p_0y)')'\cdots)'+qy=0, \tag{3}$$

where q is a continuous function of constant sign on a compact interval $[a, b]$ and $p_0, \cdots, p_n$ are positive functions such that $p_i \in C^{n-i}[a, b]$ for $i=0, \cdots, n$.

We set $L_0y=p_0y$, $L_hy=p_h(L_{h-1}y)'$ for $h=1, \cdots, n$, $Ly=L_ny$. The functions L_iy will also be called the quasi derivatives of y (see [7]).

We will be concerned with the BVP given by (3) and by a system of two-point boundary conditions of the following type (see also [2], [3], [5]):

$$L_iy(a)=0, \quad i\in\{i_1, \cdots, i_k\}, \qquad L_jy(b)=0, \quad j\in\{j_1, \cdots, j_{n-k}\}, \tag{3.1}$$

where $1\leqq k\leqq n-1$ and $i_1, \cdots, i_k, j_1, \cdots, j_{n-k}$ denote fixed integers such that $0\leqq i_1<\cdots<i_k\leqq n-1$ and $0\leqq j_1<\cdots<j_{n-k}\leqq n-1$. An important special case of system (3.1) is

$$y^{(i)}(a)=0, \quad i=0, \cdots, k-1, \qquad y^{(j)}(b)=0, \quad j=0, \cdots, n-k-1. \tag{3.2}$$

For a solution of BVP (3), (3.1) throughout the paper we always mean a nontrivial solution.

The following theorem is due to Nehari [10].

THEOREM A. *If there exists a solution of* BVP (3), (3.1) *then we have* $(-1)^{n-k}q\leqq 0$.

Nehari assumes q is strictly positive or strictly negative and considers only systems of type (3.2). But an examination of his proof shows that it is sufficient to assume only that q does not change sign and that the theorem is true also for system (3.1).

The next definition will play an essential role in the sequel.

DEFINITION. Given an index m, $0\leqq m\leqq n-1$, we will say that the system of boundary conditions (3.1) is *m-admissible* if for every integer $k=1, \cdots, n$ at least k boundary conditions are imposed on a subset of the first k terms of the sequence of quasi derivatives $L_my, L_{m+1}y, \cdots, L_{n-1}y, L_0y, \cdots, L_{m-1}y$.

The 0-admissibility is equivalent to the admissibility defined in [3] and to property (A) of [2] (see [5, Cor. 3] and [2, Thm. 1]).

Let us remark that (3.1) is m-admissible for every m, $0\leqq m\leqq n-1$, if and only if the sets $\{i_1, \cdots, i_k\}$ and $\{j_1, \cdots, j_{n-k}\}$ have empty intersection.

In the sequel we sometimes need one of the following assumptions on the function q:

(Q_1) For every $\varepsilon>0$, q does not vanish identically either in the interval $(a, a+\varepsilon)$, or in the interval $(b-\varepsilon, b)$,

(Q_2) q does not vanish identically in every subinterval of $[a, b]$.

We will also consider the eigenvalue problem (EP) given by

$$p_n(p_{n-1}\cdots(p_1(p_0y)')'\cdots)'+\lambda qy=0 \tag{3λ}$$

and the system of boundary conditions (3.1).

Eigenvalue problems of this type have been studied extensively by many authors (see [2], [3], [5], [10]). We will make use of some known results that we summarize as Theorem B.

THEOREM B. *Let us suppose that* $(-1)^{n-k}q \leqq 0$ *and* (Q_2) *holds. Then*:

(a) EP (3λ), (3.1) *has an infinite sequence* $\{\lambda_m\}_{m\in\mathbf{N}}$ *of real eigenvalues with no finite cluster points; moreover,* $\lambda_m \geqq 0$ *for every* m.

(b) *To every nonzero eigenvalue corresponds an essentially unique eigenfunction.*

(c) $\lambda = 0$ *is not an eigenvalue if and only if* (3.1) *is* 0*-admissible. Only a finite number of independent eigenfunctions correspond to* $\lambda = 0$. *If these eigenfunctions are arranged in a suitable order and the others are arranged according to the magnitude of the corresponding eigenvalue, then the* i*th eigenfunction has exactly* $i-1$ *zeros on* (a, b) *and they are all simple.*

(d) *If* $\lambda_m > 0$ *then* λ_m *is a differentiable function of the boundary points and a decreasing function of the interval* $[a, b]$ *(with respect to inclusion), moreover,* $\lambda_m \to +\infty$ *as* $b \to a^+$.

Parts (a)-(c) follow from the results of Elias [5]. Part (d) follows from Elias [6, Cor. 4] and Ahmad and Lazer [2, Lemma 3].

We also recall the results of Ahmad and Lazer [2] on the comparison with respect to changing the function q in Theorem C.

THEOREM C. *Suppose that* (3.1) *is* 0*-admissible and let* $\tilde{q} \in C[a, b]$ *be such that*

$$(-1)^{n-k}\tilde{q}(x) \leqq (-1)^{n-k}q(x) \leqq 0 \quad \text{on } [a, b] \text{ and } \tilde{q} \not\equiv q.$$

Then

(a) *If there exists a solution of* BVP (3), (3.1), *then there exists a solution of equation* $Ly + \tilde{q}(x)y = 0$ *that satisfies* (3.1) *for some* $c \in (a, b)$.

(b) *If* (Q_2) *holds and* λ_1, μ_1 *are, respectively, the first eigenvalue of* (3λ), (3.1) *and of the* EP *given by equation* $Ly + \mu\tilde{q}(x)y = 0$ *and system* (3.1), *then* $\mu_1 < \lambda_1$.

2. Positive results. In this section, using a technique inspired by Ahmad and Lazer [2], we establish comparison theorems of Sturm-Picone type for BVP (3), (3.1) and for EP (3λ), (3.1).

We begin with the case in which only one function p_m is replaced by $\tilde{p}_m$ in (3) and then we provide results for the general case. Hence we first consider the pair of BVPs

$$p_n(p_{n-1}\cdots(p_1(p_0y)')'\cdots)' + qy = 0, \tag{3}$$

$$L_iy(a) = 0, \quad i \in \{i_1, \cdots, i_k\}, \qquad L_jy(b) = 0, \quad j \in \{j_1, \cdots, j_{n-k}\}, \tag{3.1}$$

$$p_n(p_{n-1}\cdots(\tilde{p}_m\cdots(p_1(p_0u)')'\cdots)'\cdots)' + qu = 0, \tag{4}$$

$$M_iu(a) = 0, \quad i \in \{i_1, \cdots, i_k\}, \qquad M_ju(c) = 0, \quad j \in \{j_1, \cdots, j_{n-k}\}, \tag{4.1}$$

where $\tilde{p}_m$ is a function of class C^{n-m} on $[a, b]$, M_iu, $i = 0, \cdots, n$, are the quasi derivatives of u, and c is a point of $(a, b]$; and then the pair of BVPs given by (3), (3.1) and by

$$\tilde{p}_n(\tilde{p}_{n-1}\cdots(\tilde{p}_1(\tilde{p}_0v)')'\cdots)' + \tilde{q}v = 0, \tag{5}$$

$$N_iv(a) = 0, \quad i \in \{i_1, \cdots, i_k\}, \qquad N_jv(c) = 0, \quad j \in \{j_1, \cdots, j_{n-k}\}, \tag{5.1}$$

where $\tilde{p}_i \in C^{n-i}[a, b]$, $\tilde{q} \in C[a, b]$ and N_iv, $i = 0, \cdots, n$, are the quasi derivatives of v.

The main result of this section is the following theorem.

THEOREM 1. *Suppose that* (3.1) *is* m*-admissible,* $0 \leqq m \leqq n-1$, *and that*

$$0 < \tilde{p}_m(x) \leqq p_m(x), \qquad x \in [a, b].$$

If there exists a solution of (3), (3.1) *then there exists a solution of* (4), (4.1) *for some* $c \leqq b$. *Moreover, if* $\tilde{p}_m \not\equiv p_m$ *and* (Q_1) *holds then* c *can be chosen less than* b.

In § 3 we show that the hypothesis of m-admissibility cannot be dropped; we will also show that if (Q_1) does not hold, then (4), (4.1) can have no solution for every $c \in (a, b)$ even if $\tilde{p}_m \not\equiv p_m$.

If $\{i_1, \cdots, i_k\} \cap \{j_1, \cdots, j_{n-k}\} = \varnothing$, then the system (3.1) is m-admissible for every m, hence by Theorems 1 and C we deduce the following theorem.

THEOREM 2. *Suppose that*

$$0 < \tilde{p}_i(x) \leqq p_i(x), \quad x \in [a, b], \quad i = 0, \cdots, n,$$

$$(-1)^{n-k}\tilde{q}(x) \leqq (-1)^{n-k}q(x) \leqq 0, \qquad x \in [a, b],$$

$$\{i_1, \cdots, i_k\} \cap \{j_1, \cdots, j_{n-k}\} = \varnothing,$$

$$\tilde{q} \not\equiv q \text{ or else } \tilde{p}_i \not\equiv p_i \text{ for at least one index } i \text{ and } (Q_1) \text{ holds.}$$

If there exists a solution of (3), (3.1), *then there also exists a solution of* (5), (5.1) *for some* $c \in (a, b)$.

Corresponding results can be stated also for the eigenvalue problems given, respectively, by

$$(3\lambda) \qquad p_n(p_{n-1} \cdots (p_1(p_0 y)')' \cdots)' + \lambda q y = 0$$

and system (3.1); by

$$(4\lambda) \qquad p_n(p_{n-1} \cdots (\tilde{p}_m \cdots (p_1(p_0 u)')' \cdots)' \cdots)' + \mu q u = 0$$

and system (4.1); and by

$$(5\lambda) \qquad \tilde{p}_n(\tilde{p}_{n-1} \cdots (\tilde{p}_1(\tilde{p}_0 v)')' \cdots)' + \sigma \tilde{q} v = 0$$

and system (5.1).

THEOREM 3. *Suppose that* (3.1) *is m-admissible,* $0 \leqq m \leqq n-1$, *and that* $(-1)^{n-k}q \leqq 0$, (Q_2) *holds, and*

$$0 < \tilde{p}_m(x) \leqq p_m(x), \quad x \in [a, b], \quad \tilde{p}_m \not\equiv p_m.$$

If λ_0 *and* μ_0 *are, respectively, the first positive eigenvalue of* EP (3λ), (3.1) *and of* EP (4λ), (4.1), *then* $\mu_0 < \lambda_0$.

From Theorems 3 and C we have Theorem 4.

THEOREM 4. *Suppose that* (Q_2) *holds and that*

$$0 < \tilde{p}_i(x) \leqq p_i(x), \quad x \in [a, b], \quad i = 0, \cdots, n,$$

$$(-1)^{n-k}\tilde{q}(x) \leqq (-1)^{n-k}q(x) \leqq 0, \qquad x \in [a, b],$$

$$\{i_1, \cdots, i_k\} \cap \{j_1, \cdots, j_{n-k}\} = \varnothing,$$

$$\tilde{q} \not\equiv q \text{ or else } \tilde{p}_i \not\equiv p_i \text{ for at least one index } i.$$

If λ_0 *and* σ_0 *are, respectively, the first positive eigenvalue of* (3λ), (3.1) *and of* (5λ), (5.1), *then* $\sigma_0 < \lambda_0$.

The proofs of Theorems 1 and 3 follow the pattern of the proof of Theorem C (see [2]). At first we prove a lemma (Lemma 3) which, if $q(x) \not\equiv 0$ on $[a, b]$, is an extension of Lemma 4 of [2], then Theorem 3 and finally Theorem 1. We also use the same notation as [2].

Let f be a continuous function defined on $[a, b]$. A maximal closed subinterval of $[a, b]$, which may consist of a single point, on which f is identically zero, will be called a *zero component* of f. A zero component of f will be called *internal* if it does not contain the points a and b. We say that f *changes sign* h times on a subinterval of

$[a, b]$ if there exist $h+1$ points $x_1 < \cdots < x_{h+1}$ belonging to this interval such that $f(x_i)f(x_{i+1}) < 0$ for $i = 1, \cdots, h$. We say that f changes sign on a subinterval if f changes sign at least once on the subinterval.

LEMMA 1. *Let $f_1, \cdots, f_{m+1}$, be continuous functions on $[a, b]$ and suppose that for $1 \leqq h \leqq m$ the function f_h does not vanish identically on $[a, b]$. Suppose also that for all h, $1 \leqq h \leqq m$, f_{h+1} changes sign in the open interval between two zero components of f_h. If $m+s$, $s \geqq 0$, of the numbers $f_1(a), \cdots, f_m(a), f_1(b), \cdots, f_m(b)$ are zero, and f_1 has t, $t \geqq 0$, internal zero components, then f_{m+1} changes sign at least $s+t$ times on $[a, b]$.*

(When $t = 0$, Lemma 1 is exactly part (I) of Lemma 1 of [2].)

Proof. We prove the lemma by induction on m.

For $m = 1$ the claim is trivial. Assume that the lemma is true for m or less functions. If $s = 0$ and one of the numbers $f_m(a), f_m(b)$, is nonzero or if $s \geqq 1$, then by applying the induction hypothesis to $f_1, \cdots, f_{m-1}, f_m$ we have trivially that f_{m+1} changes sign at least $s+t$ times. Suppose now that $s = 0$ and $f_m(a) = f_m(b) = 0$. If $f_1(a) \neq 0$ and $f_1(b) \neq 0$ then f_2 has at least $t-1$ internal zero components. By applying the induction hypothesis to $f_2, \cdots, f_{m+1}$ we have the claim. Otherwise we denote by k the greatest index such that exactly k of the numbers $f_1(a), \cdots, f_k(a), f_1(b), \cdots, f_k(b)$ are zero. By applying the induction hypothesis to $f_1, \cdots, f_{k+1}$, we have that f_{k+1} changes sign at least t times, and hence f_{k+2} changes sign at least $t-1$ times. By definition of k we have $f_{k+1}(a) \neq 0$, $f_{k+1}(b) \neq 0$. Therefore $m-k$ of the numbers $f_{k+2}(a), \cdots, f_m(a)$, $f_{k+2}(b), \cdots, f_m(b)$ are equal to zero. By applying the induction to $f_{k+2}, \cdots, f_{m+1}$, we have the claim. □

We denote by $N(i)$, $i = 0, \cdots, n-1$, the number of indexes of the set $\{i_1, \cdots, i_k, j_1, \cdots, j_{n-k}\}$ that are equal to i in system (3.1). Let

$$Z(m) = \sum_{i=m}^{n-1} (N(i) - 1), \qquad m = 0, \cdots, n-1,$$

$$Z = \max \{Z(m): m = 0, \cdots, n-1\}.$$

We have $Z(0) = 0$; hence $Z \geqq 0$.

Simple criteria to test the m-admissibility are given in the next lemma.

LEMMA 2. (a) *System* (3.1) *is m-admissible if and only if $Z(m) = Z$.*

(b) *Suppose that $q \not\equiv 0$. If there exists a solution y of* (3), (3.1) *such that $L_m y(x) \geqq 0$ on $[a, b]$ then system* (3.1) *is m-admissible.*

(c) *If* (Q_2) *holds, $(-1)^{n-k} q(x) \leqq 0$ on $[a, b]$ and y is one eigenfunction corresponding to the first positive eigenvalue of* (3λ), (3.1), *then y has exactly Z zeros on (a, b). Moreover, $L_m y(x) \neq 0$ on (a, b) if and only if system* (3.1) *is m-admissible.*

Proof. (a) Suppose that $Z(m) < Z$ and let s be an index such that $Z(s) = Z$. If $s > m$ then over the $s-m$ quasi derivatives $L_m y, \cdots, L_{s-1} y$ are imposed less than $s-m$ boundary conditions, hence (3.1) is not m-admissible. If $s < m$ then over the $m-s$ quasi derivatives $L_s y, \cdots, L_{m-1} y$ are imposed more than $m-s$ conditions. Therefore on the $n-m+s$ quasi derivatives $L_m y, \cdots, L_{n-1} y$, $L_0 y, \cdots, L_{s-1} y$ are imposed less than $n-m+s$ boundary conditions, hence (3.1) is not m-admissible.

In the same way we can prove the converse.

(b) Let us consider the $n+1$ functions $f_1, \cdots, f_{n+1}$ defined by $f_i = L_{m+i-1} y$ for $i = 1, \cdots, n-m$, $f_i = L_{i-n+m-1} y$ for $i = n-m+1, \cdots, n+1$. If for some index i, $f_i \equiv 0$, then $Ly \equiv 0$, hence $qy \equiv 0$. Since y cannot vanish in a subinterval of $[a, b]$ it follows that $q \equiv 0$, contrary to the assumptions. Since q does not change sign, as a consequence of Rolle's theorem it follows that for $i = 1, \cdots, n$, f_{i+1} changes sign between two zero components of f_i. Hence $f_1, \cdots, f_{n+1}$ satisfy the hypotheses of Lemma 1. If system

(3.1) is not m-admissible then there exists an index s such that $n-s+2$ out of the numbers $f_s(a), \cdots, f_n(a), f_s(b), \cdots, f_n(b)$ are zero. By Lemma 1 we have that $f_{n+1} = L_m y$ changes sign on $[a, b]$, contrary to the assumptions.

(c) If $Z = 0$ then system (3.1) is 0-admissible, therefore by Theorem B(c), y does not vanish on (a, b). Suppose that $Z \geqq 1$ and let s be the lowest index such that $Z(s) = Z$. Among the numbers $L_s y(a), \cdots, L_{n-1} y(a), L_s y(b), \cdots, L_{n-1} y(b)$ $n-s+Z$ are equal to zero. By Lemma 1, if Z_1 is the number of zeros of y, then $Z_1 \geqq Z$. Let us suppose that $Z_1 > Z$. By Theorem B(c) there exists one eigenfunction w, corresponding to $\lambda = 0$, with Z zeros on (a, b). Since $L_n w \equiv 0$, a repeated integration shows that there exists an index $t \leqq n-1$ such that $L_t w$ does not vanish on $[a, b]$.

Suppose that $t \geqq s$. By definition of s, at least $t-s+1$ of the numbers $L_s w(a), \cdots, L_t w(a), L_s w(b), \cdots, L_t w(b)$ are equal to zero. Since $L_t w$ does not vanish, we have that $t > s$ and that $t-s+1$ out of the numbers $L_s w(a), \cdots, L_{t-1} w(a)$, $L_s w(b), \cdots, L_{t-1}(b)$ are equal to zero. Applying Lemma 1 to the functions $L_s w, \cdots, L_{t-1} w, L_t w$, we have that $L_t w$ changes sign, in contradiction to the definition of t.

Suppose that $t < s$. By definition of s, fewer than $s-t$ boundary conditions are imposed on $L_t y, \cdots, L_{s-1} y$; therefore more than $t-Z$ boundary conditions are imposed on $L_0 y, \cdots, L_{t-1} y$. By Lemma 1 we again have that $L_t w$ changes sign. This contradiction shows that $Z_1 = Z$.

By (b), if $L_m y(x) \neq 0$ on (a, b), then (3.1) is m-admissible. Conversely, if (3.1) is m-admissible and $L_m y$ has a zero on (a, b) then by Lemma 1, y changes sign $Z+1$ times, contradicting what we have already proved. □

Example. Let $n = 6$ and consider the following system of boundary conditions.

$$L_0 y(a) = L_2 y(a) = L_3 y(a) = L_5 y(a) = 0,$$

$$L_3 y(b) = L_5 y(b) = 0.$$

We have $Z(5) = 1$, $Z(4) = 0$, $Z(3) = 1$, $Z(2) = 1$, $Z(1) = 0$, $Z(0) = 0$; hence $Z = 1$. By Lemma 2 this system of boundary conditions is m-admissible for $m = 2, 3, 5$, whereas for the other m is not m-admissible.

LEMMA 3. *Let us suppose that* $q \not\equiv 0$ and that

$$0 < \tilde{p}_m(x) \leqq p_m(x), \qquad x \in [a, b]. \tag{6}$$

Assume that there exists a solution y *of* (3), (3.1) *and a solution* u *of* (4), (4.1) *where* $c = b$. *If* $M_m u(x) \geqq 0$ *on* $[a, b]$ *then there exists* $\lambda \in \mathbf{R}$ *and an interval* $[x_1, x_2] \subset [a, b]$, $x_1 < x_2$ *such that* $M_m u(x) = \lambda L_m y(x)$ *on* $[a, b]$ *and* $p_m(x) = \tilde{p}_m(x)$ *on* $[x_1, x_2]$. *Moreover, if* (Q_1) *holds, then* $u \equiv \lambda y$ *and* $\tilde{p}_m \equiv p_m$.

Proof. Step 1. For every $\gamma \in \mathbf{R}$ let us consider the function

$$w_\gamma = M_m u + \gamma L_m y.$$

We want to prove the following assertion.

Let $\{\gamma_n\}$ be a sequence of positive real numbers converging to γ_0. If w_{γ_n} changes sign on $[a, b]$ for every $n \geqq 1$ and $w_{\gamma_0}(x) \geqq 0$ on $[a, b]$, then $w_{\gamma_0} \equiv 0$.

For all integer $s \geqq 0$ we consider $n+1$ functions $f_{i,s}(x)$, $x \in [a, b]$ defined in the following way

$$f_{i,s} = M_{m-1+i} u + \gamma_s L_{m-1+i} y \quad \text{for } i = 1, \cdots, n-m,$$

$$f_{i,s} = L_{m-1-n+i}(u + \gamma_s y) \quad \text{for } i = n-m+1, \cdots, n+1.$$

Since q does not change sign on $[a, b]$, Rolle's theorem implies that for all $s \geqq 0$ and

for $i=1,\cdots,n$, $f_{i+1,s}$ changes sign between two zero components of $f_{i,s}$. Moreover, since y and u satisfy, respectively, systems (3.1) and (4.1), for all $s\geqq 0$, n of the numbers $f_{1,s}(a),\cdots,f_{n,s}(a), f_{1,s}(b),\cdots,f_{n,s}(b)$ are zero.

Since $Mu\geqq 0$ on $[a,b]$, by Lemma 2 systems (3.1) and (4.1) are m-admissible. By hypothesis for all $s\geqq 1$, $f_{1,s}$ changes sign on $[a,b]$. By the definition of m-admissibility and by Lemma 1, $f_{i,s}$ changes sign on $[a,b]$ for every i. Therefore for every i and for every $s\geqq 1$, $f_{i,s}$ does not vanish identically on $[a,b]$.

Again by the m-admissibility and Lemma 1, it follows that for every $i\leqq n-m$, $f_{i,0}$ has a zero on $[a,b]$. If for an index $i\leqq n-m$, $f_{i,0}\equiv 0$, then $w_{\gamma_0}=f_{1,0}\equiv 0$. If for an index $i\geqq n-m+1$, $f_{i,0}\equiv 0$ then $f_{n+1,0}\equiv 0$. But by (6) we have

$$f_{n+1,0}\equiv L_m u+\gamma_0 L_m y\geqq M_m u+\gamma_0 L_m y\geqq 0, \tag{7}$$

and therefore $w_{\gamma_0}\equiv 0$.

Hence either $w_{\gamma_0}\equiv 0$ or the functions $f_{i,s}$, $i=1,\cdots,n+1$ satisfy the hypotheses of Lemma 1 for all $s\geqq 0$. We now show that in the latter case we have a contradiction.

At first we prove that $f_{n,0}$ vanishes at one of the endpoints. Assume the contrary. It follows that in system (3.1) no boundary conditions are imposed on $L_{m-1}y$ ($L_{n-1}y$ if $m=0$). By Lemma 1, for every $s\geqq 1$, $f_{n,s}$ has two distinct zero components on $[a,b]$. Since $f_{n,s}$ converges to $f_{n,0}$ and since, by (7), $f_{n,0}$ cannot have two distinct zero components, there exists a unique maximal closed subinterval $[x_1,x_2]$, $a<x_1\leqq x_2<b$, of $[a,b]$, on which $f_{n,0}$ is identically zero. Between two zero components of $f_{n,0}$ there exists a zero component of $f_{n+1,0}$; hence on $[x_1,x_2]$ we have also $f_{n+1,0}(x)=0$. By (7), it follows that $f_{1,0}(x)=0$ on $[x_1,x_2]$. If $f_{1,0}$ has an internal zero component, then, by Lemma 1, $f_{n+1,0}$ changes sign in contradiction to (7), and hence $f_{1,0}$ vanishes identically on one of the intervals $[a,x_1]$ or $[x_2,b]$. Without loss of generality we may assume that $f_{1,0}$ vanishes identically on $[a,x_1]$. If there exists $t_1\in(a,x_1)$ such that $q(t_1)\neq 0$, then there exists an interval $[t_1,t_2]$, $t_1<t_2$, where $u+\gamma_0 y\equiv 0$. Therefore we have $f_{n,0}(t_1)=0$, in contradiction to the definition of x_1. Hence $q(x)=0$ on $[a,x_1]$. If $\{m,\cdots,n-1\}\not\subset\{i_1,\cdots,i_k\}$, by Lemma 1 $f_{n+1,0}$ changes sign on $[a,b]$, hence $M_i u(a)=0$ for $i=m,\cdots,n-1$. So $M_m u(x)=0$ on $[a,x_1]$ and since $\tilde{p}_m>0$ we also have that $L_m u(x)=0$ on $[a,x_1]$. Therefore $f_{n+1,0}(x)=0$ on $[a,x_1]$. Hence $f'_{n,0}(x)=0$ on $[a,x_1]$, in contradiction to the definition of x_1. Hence either $f_{n,0}(a)=0$ or $f_{n,0}(b)=0$.

Suppose that $f_{n,0}(a)=0$. Let i, $2\leqq i\leqq n$, be an index such that $f_{i,0}(a)=f_{i+1,0}(a)=\cdots=f_{n,0}(a)=0$. If for some index j, $2\leqq j\leqq n$, $f_{j,0}(b)=0$, then, by Lemma 1, $f_{n+1,0}$ changes sign, in contradiction to (7). It follows that no boundary conditions are imposed in system (3.1) on $L_i y(b),\cdots,L_n y(b)$. If there are no boundary conditions imposed on $L_{i-1}y$, then, by Lemma 1, $f_{i-1,s}$ changes sign twice on $[a,b]$. If there is only one boundary condition imposed on $L_{i-1}y$, then, by Lemma 1, $f_{i-1,s}$ changes sign at least once on $[a,b]$. Therefore, in any case, $f_{i-1,s}$ has two zero components on $[a,b]$. If $f_{i-1,0}$ has two zero components on $[a,b]$ then, by Lemma 1, $f_{n+1,0}$ changes sign. Since between two zero components of $f_{i-1,0}$ there exists a zero component of $f_{i,0}$, we have that $f_{i-1,0}$ and $f_{i,0}$ have a common zero component on $[a,b]$. Since $f_{i,0}$ cannot have two distinct zero components, it follows that $f_{i-1,0}(a)=0$. So if $f_{n,0}(a)=0$, then $f_{1,0}(a)=\cdots=f_{n,0}(a)=0$. In the same way if $f_{n,0}(b)=0$, then $f_{1,0}(b)=\cdots=f_{n,0}(b)=0$. Since there is at least one boundary condition imposed on points a and b, in both cases, by Lemma 1, $f_{n+1,0}$ changes sign in contradiction to (7).

Step 2. If $M_m u\equiv 0$, then $qu\equiv 0$. Since a solution of (3) cannot vanish identically in a subinterval of $[a,b]$, we have $q\equiv 0$, contrary to the assumptions. Hence there exists $t_0\in[a,b]$ such that $M_m u(t_0)>0$. By the same argument there exists $t_1\in[a,b]$ such that $L_m y(t_1)\neq 0$. Interchanging possibly y by $-y$, we may assume that $L_m y(t_1)<0$.

Let E be the set of all positive numbers γ such that $w_\gamma(x) \geqq 0$ on $[a, b]$. E is not empty, in fact otherwise, since $M_m u(t_0) > 0$, there exists $\delta > 0$ such that for $0 < \gamma < \delta$, w_γ changes sign; by Step 1 we have $M_m u \equiv 0$, which is a contradiction. Moreover, E is bounded above because $L_m y(t_1) < 0$. Let $\lambda = \sup E$. Obviously, $w_\lambda(x) \geqq 0$ on $[a, b]$. If there exists $\delta > 0$ such that for $\lambda < \gamma < \lambda + \delta_1 w_\gamma(x) \leqq 0$ on $[a, b]$, then $w_\lambda \equiv 0$; otherwise by applying Step 1 we have again $w_\lambda \equiv 0$. Hence $M_m u \equiv -\lambda L_m y$ and $q(u + \lambda y) \equiv 0$. Since $q \not\equiv 0$ there exists an interval $[x_1, x_2]$ where $q(x) \neq 0$. In this interval we have $u + \lambda y \equiv 0$ and

$$\tilde{p}_m(L_{m-1}u)' + \lambda p_m(L_{m-1}y)' \equiv (\tilde{p}_m - p_m)(L_{m-1}u)' \equiv 0. \tag{8}$$

If $(L_{m-1}u)'$ vanishes identically on a subinterval of $[x_1, x_2]$ then on this subinterval we have $u \equiv 0$ in contradiction to the definition of u. Hence by (8) it follows that $\tilde{p}_m(x) = p_m(x)$ on $[x_1, x_2]$.

Finally, if (Q_1) holds, then $L_i u(a) = -\lambda L_i y(a)$ and $L_i u(b) = -\lambda L_i y(b)$ for $i = 0, \cdots, m-1$. Hence if $u \not\equiv \lambda y$ then, by Lemma 1, $L_m(u + \lambda y)$ changes sign in contradiction to (7). Moreover, since (Q_1) holds, $M_m u$ cannot vanish in a neighborhood of a and of b, and by Lemma 1, $M_m u$ cannot have a zero component that does not contain the point a or b; hence $M_m u(x) > 0$ on (a, b) and by (8) we have $\tilde{p}_m \equiv p_m$. □

Proof of Theorem 3. Let y be one eigenfunction corresponding to λ_0 and let u be one eigenfunction corresponding to μ_0. By assumption system (3.1) and system (4.1) are m-admissible, therefore by Lemma 2, possibly interchanging u by $-u$, we have that $M_m u(x) > 0$ and (a, b). If $\mu_0 \geqq \lambda_0$ then $0 < \tilde{p}_m(x)/\mu_0 \leqq p_m(x)/\lambda_0$ on $[a, b]$. By Lemma 3 it follows that $\lambda_0 \tilde{p}_m \equiv \mu_0 p_m$ and then $\tilde{p}_m \geqq p_m$, contrary to the assumptions. □

Remark. Ahmad and Lazer have shown in [2] (see the proof of Theorem 1 and of Lemma 3), that BVP (3), (3.1) has the following property. Let $\{c_m\}$ be a sequence of points of $(a, b]$ converging to c and let $\{q_m\}$ be a sequence of continuous nonvanishing functions converging uniformly on $[a, b]$ to q. If for every m the equation $Ly + q_m(x)y = 0$ has a solution y_m satisfying (3.1) where $b = c_m$, then $c > a$ and there also exists a solution y of equation $Ly + q(x)y = 0$ that satisfies (3.1) with $b = c$. Moreover, multiplying every function y_m by a suitable constant, considering y_m defined on the entire interval $[a, b]$, and considering a suitable subsequence, we may assume that $\{L_i y_m\}$ converges uniformly to $L_i y$ on $[a, c]$, for $i = 0, \cdots, n-1$.

Proof of Theorem 1. If $q \equiv 0$ the theorem is a trivial consequence of Theorem B. Since q is a continuous function, we can suppose then that there exists an interval $[x_1, x_2] \subset [a, b]$, where q does not vanish. Since BVP (3), (3.1) has a nontrivial solution, by Theorem A we have that $(-1)^{n-k} q(x) \leqq 0$ on $[a, b]$.

Let us consider the sequence of functions $\{q_s\}$ defined as follows.

$$q_s(x) = \begin{cases} 0 & \text{if } x_1 < x < x_2, \\ (-1)^{n-k}(x - x_1)/s & \text{if } a \leqq x \leqq x_1, \\ (-1)^{n-k}(x_2 - x)/s & \text{if } x_2 \leqq x \leqq b. \end{cases}$$

Let $\lambda_{0,s}$ be the first positive eigenvalue of EP given by

$$L_n y + \lambda(q_s(x) + q(x))y = 0$$

and system of boundary conditions (3.1) and let y_s be one eigenfunction corresponding to $\lambda_{0,s}$. If λ_0 is the first positive eigenvalue of (3λ), (3.1), where $a = x_1$ and $b = x_2$, then by Theorem B(d) we have $\lambda_{0,s} \leqq \lambda_0$ for every s. Hence the sequence $\{\lambda_{0,s}\}$ is bounded. By Lemma 2, $L_m y_s(x)$ does not vanish on (a, b). Considering a suitable subsequence and by replacing y_s by $-y_s$ if necessary, we may assume that $\lambda_{0,s} \to \lambda^*$ and $L_m y_s(x) > 0$

on (a, b). By the previous remark we have that there also exists a solution w of BVP given by

$$L_n y + \lambda^* q(x) y = 0$$

and system (3.1), such that $L_i y_s$ converges uniformly to $L_i w$ for every i, $0 \leqq i \leqq n-1$. Since $L_m y_s(x) > 0$ on (a, b) we have $L_m w(x) \geqq 0$ on $[a, b]$.

If $\lambda^* > 1$, then $p_m/\lambda^* < p_m$. Since w is a solution of BVP (4), (4.1) with $\tilde{p}_m = p_m/\lambda^*$, satisfying $L_m w(x) \geqq 0$ on $[a, b]$ we have a contradiction to Lemma 3. Hence $\lambda^* \leqq 1$.

Let $\mu_{0,s}$ be the first positive eigenvalue of EP given by

$$M_n u + \mu(q_s(x) + q(x))u = 0 \tag{9}$$

and system (4.1) and let u_s be an eigenfunction corresponding to $\mu_{0,s}$. By Theorem 3, $\mu_{0,s} < \lambda_{0,s}$. Considering a suitable subsequence we may assume that $\mu_{0,s} \to \mu^* \leqq \lambda^*$. If $\mu^* = 1$ then by the previous remark there exists a solution of (4), (4.1) where $c = b$. Suppose that $\mu^* < 1$. Considering a suitable subsequence we may assume $\mu_{0,s} < 1$ for all s. By Theorem B(d) there exists $c_s \in (a, b)$ such that the first positive eigenvalue of (9), (4.1), where $c = c_s$, is 1. The sequence $\{c_s\}$ has a subsequence converging to a point $c \in [a, b]$. By the previous remark it follows that $c > a$ and that (4), (4.1) has a solution.

Suppose now that $\tilde{p}_m \not\equiv p_m$ and that (Q_1) holds.

Let us denote by u the solution of (4), (4.1). By Lemma 2 every function $M_m u_s$ does not change sign on $[a, b]$. Again by the previous remark it follows that also $M_m u$ does not change sign on $[a, b]$. Interchanging u and $-u$ if necessary, we have that $M_m u \geqq 0$ on $[a, b]$. If $c = b$ by Lemma 3 we have $\tilde{p}_m \equiv p_m$, contrary to the assumptions. □

3. Negative results. The aim of this section is to show that when only one boundary condition is imposed at one of the endpoints and $n \geqq 3$, then the m-admissibility of system (3.1) is a necessary and sufficient condition for the validity of Theorems 1 and 3. Clearly, when $n = 3$ we are always in such a situation. Hence the extension of the Sturm-Picone theorem to (1) or (2) is true if and only if the system of boundary conditions is 1-admissible in the case of (1), or 2-admissible in the case of (2). Since system (1.1) is neither 1-admissible nor 2-admissible, it follows that Lazer's question has a negative answer in both cases.

From Theorems 5 and 6 below, it follows that for third-order equations the general comparison, as given in Theorems 2 and 4, is true if and only if exactly one boundary condition is imposed on every quasi derivative of y (i.e., $\{i_1, \cdots, i_k\} \cap \{j_1, \cdots, j_{n-k}\} = \varnothing$).

We will consider the equation $y^{(n)} + q_0 y = 0$, where q_0 is a nonvanishing constant. In this case problems (3), (3.1) and (4), (4.1) become

$$y^{(n)} + q_0 y = 0, \tag{10}$$

$$y^{(i_1)}(a) = \cdots = y^{(i_k)}(a) = 0, \qquad y^{(j_1)}(b) = \cdots = y^{(j_{n-k})}(b) = 0, \tag{10.1}$$

$$(\tilde{p}_m u^{(m)})^{(n-m)} + q_0 u = 0, \tag{11}$$

$$M_i u(a) = 0, \quad i \in \{i_1, \cdots, i_k\}, \qquad M_j u(c) = 0, \quad j \in \{j_1, \cdots, j_{n-k}\}, \tag{11.1}$$

where $M_i u = u^{(i)}$ for $i = 0, \cdots, m-1$; $M_i u = (\tilde{p}_m u^{(m)})^{(i-m)}$ for $i = m, \cdots, n$.

As in the previous section we will also be concerned with the corresponding eigenvalue problems given, respectively, by

$$y^{(n)} + \lambda q_0 y = 0 \tag{10λ}$$

and system (10.1) and by

$$(11\lambda) \qquad (\tilde{p}_m u^{(m)})^{(n-m)} + \mu q_0 u = 0$$

and system (11.1).

The principal negative results are given in Theorems 5 and 6.

THEOREM 5. *Assume that* $n \geqq 3$, *that* $k = 1$ *or* $k = n-1$ *and that there exists a solution of* (10), (10.1).

If system (10.1) *is not m-admissible,* $0 \leqq m \leqq n-1$, *then there exist a point* $b_1 \in (a, b]$ *and a function* $\tilde{p}_m \in C^{n-m}[a, b]$ *such that a solution of* (10) *exists on* $[a, b_1]$ *that satisfies the boundary conditions* (10.1) *with* $b = b_1$; $0 < \tilde{p}_m(x) \leqq 1$ *on* $[a, b]$; *for every* $c \in (a, b_1]$, BVP (11), (11.1) *has no solutions.*

THEOREM 6. *Assume that* $n \geqq 3$, *that* $k = 1$ *or* $k = n-1$, *and that* $(-1)^{n-k} q_0 < 0$.

If system (10.1) *is not m-admissible* $0 \leqq m \leqq n-1$, *then there exists a function* $\tilde{p}_m \in C^{n-m}[a, b]$ *such that* $0 < \tilde{p}_m(x) \leqq 1$ *on* $[a, b]$ *and the first positive eigenvalue of* (11λ), (11.1) *is greater than the first positive eigenvalue of* (10λ), (10.1).

Theorem 6 shows that Theorem 1.3 of Gentry and Travis [9, p. 169] is not correct. In fact it states, in the case of (10), that if we consider the system of boundary conditions of type (3.2) (which is m-admissible only for $m = 0$) and take $m = n-1$, such a function $\tilde{p}_m$ cannot exist.

To prove Theorems 5 and 6 we need some results established by the author in [8]. Hence we recall some definitions used there.

We say that a nonnull vector η of $\mathbf{R}^n$, $\eta = (\eta_1, \cdots, \eta_n)$, has the D-property if three indices i, j, k do not exist such that $1 \leqq i < j < k \leqq n$ and $\eta_i \eta_j < 0$, $\eta_j \eta_k < 0$.

We say that η has the strict D-property in the case where $\eta_i \eta_j > 0$ for all $i, j \geqq 2$ or $\eta_i \eta_j > 0$ for all $i, j \leqq n-1$ or else there exists an index k, $1 < k < n$, such that for all $i < k$ and $j > k$, $\eta_i \eta_j < 0$.

If η has the D-property, we indicate with $r(\eta)$ the greatest index j such that $\eta_j \neq 0$ and $\eta_i \eta_j \geqq 0$ for every $i \leqq j$.

Let us consider the Cauchy problem (CP).

$$(12) \qquad y^{(n)} + q(x) y = 0,$$

$$(12.1) \qquad y^{(i)}(\xi) = \eta_{i+1}, \qquad i = 0, \cdots, n-1,$$

where $q(x) > 0$. Let $e_1, \cdots, e_n$ be the canonical basis of $\mathbf{R}^n$. If $\eta = e_s$ for a given s, the solution of (12), (12.1) will be denoted by u_s and will be called a principal solution of (12).

By Rolle's theorem, the derivatives of a nontrivial solution of (12), (12.1) have a finite number of zeros on an interval $(\xi, c]$, $c > \xi$. Let $z_1, \cdots, z_m$ be the ordered set (possibly empty) of these zeros. In [8] we proved the following lemma.

LEMMA 4. *If* η *has the D-property and* y *is the solution of* (12), (12.1), *then the vector* $Y(x) = (y(x), y'(x), \cdots, y^{(n-1)}(x))$ *has the strict D-property for every* $x \in (\xi, c]$. *Moreover,* $y^{(j)}$, $0 \leqq j \leqq n-1$, *vanish at the point* z_i, $i \geqq 1$, *if and only if* $j \equiv (r(\eta) - i) \bmod n$.

The next lemma is derived from the proposition of [8, p. 240] and can be proved using the same argument.

LEMMA 5. *Suppose that* $u_s^{(j)}$, $1 \leqq s \leqq n$, $0 \leqq j \leqq n-1$, *has m zeros* $w_1, \cdots, w_m$ *on* $(\xi, c]$. *If* η *has the D-property,* $r(\eta) = s$, *and there exists an index* $i \neq s$ *such that* $\eta_i \neq 0$, *then the j-derivative of the solution of* (12), (12.1) *has exactly m zeros* $z_1, \cdots, z_m$ *on* $(\xi, w_m]$ *and we have* $z_i < w_i$ *for every i.*

Let m, $1 \leqq m \leqq n-1$, be a given index, let $k \in \mathbf{R}^+$, and let t_1, t_2 be two points of $(a, +\infty)$. We consider the functions z_s, $1 \leqq s \leqq n$, defined as follows.

If $x \in [a, t_1]$ then z_s is the solution of CP,

$$y^{(n)} + q_0 y = 0, \tag{10}$$

$$y^{(i)}(a) = 0, \quad i = 0, 1, \cdots, s-2, s, \cdots, n-1, \quad y^{(s-1)}(a) = 1. \tag{10.2}$$

If $x \in (t_1, t_2)$ then z_s is the solution of CP,

$$y^{(n)} + \frac{q_0}{k} y = 0, \tag{13}$$

$$y^{(i)}(t_1) = z_s^{(i)}(t_1), \quad i = 0, \cdots, m-1, \qquad y^{(i)}(t_1) = \frac{z_s^{(i)}(t_1)}{k}, \quad i = m, \cdots, n-1. \tag{13.1}$$

If $x \geqq t_2$ then z_s is the solution of CP,

$$y^{(n)} + q_0 y = 0, \tag{10}$$

$$y^{(i)}(t_2) = \lim_{x \to t_2^-} z_s^{(i)}(x), \quad i = 0, \cdots, m-1, \qquad y^{(i)}(t_2) = k \lim_{x \to t_2^-} z_s^{(i)}(x), \quad i = m, \cdots, n-1. \tag{10.3}$$

LEMMA 6. *Let $0 < k \leqq 1$, $1 \leqq s \leqq n$, and $a < t_1 < t_2 < b$. For every $\varepsilon > 0$ there exists $r \in C^\infty[a, b]$ such that $k \leqq r(x) \leqq 1$ on $[t_1, t_2]$, $r(x) = 1$ on $[a, t_1] \cup [t_2, b]$ and the principal solution u_s of $(ry^{(m)})^{(n-m)} + q_0 y = 0$ is such that*

$$|u_s^{(i)}(x) - z_s^{(i)}(x)| \leqq \varepsilon \quad \text{for } x \in [a, b] \quad \text{and} \quad i = 0, \cdots, m-1,$$

$$|u_s^{(i)}(x) - z_s^{(i)}(x)| \leqq \varepsilon \quad \text{for } x \in [a, t_1] \cup [t_2, b] \quad \text{and} \quad i = m, \cdots, n-1,$$

$$|(r(x)u_s^{(m)}(x))^{(i-m)} - k z_s^{(i)}(x)| \leqq \varepsilon \quad \text{for } x \in (t_1, t_2) \quad \text{and} \quad i = m, \cdots, n-1.$$

Proof. Let us consider for every δ, $0 < \delta < 1$, the function

$$r_\delta(x) = \begin{cases} 1 & \text{if } x \leqq t_1 \text{ or } x \geqq t_2, \\ 1 - (1-k) \exp\left(\dfrac{(t_2 - t_1)^2 \log(1-\delta)}{4(x-t_1)(t_2-x)} \right) & \text{if } t_1 < x < t_2. \end{cases}$$

We remark that the CP

$$(r_\delta y^{(m)})^{(n-m)} + q_0 y = 0, \qquad y^{(i)}(t_1) = z_s^{(i)}(t_1), \quad i = 0, \cdots, n-1$$

is equivalent to the CP given by the system

$$y_h' = y_{h+1}, \qquad h = 1, \cdots, m-1, m+1, \cdots, n-1,$$

$$y_m' = \frac{y_{m+1}}{r_\delta},$$

$$y_n' = -q_0 y_1,$$

and initial conditions $y_i(t_1) = z_s^{(i-1)}(t_1)$, $i = 1, \cdots, n$, whereas the CP (13), (13.1) is equivalent to the CP given by the system

$$w_h' = w_{h+1}, \qquad h = 1, \cdots, m-1, m+1, \cdots, n-1,$$

$$w_m' = \frac{w_{m+1}}{k},$$

$$w_n' = -q_0 w_1,$$

and initial conditions $w_i(t_1) = z_s^{(i-1)}(t_1)$, $i = 1, \cdots, n$.

Hence, since $k \leqq r_\delta(x) \leqq 1$ on $[t_1, t_2]$ and as $\delta \to 0^+$, $r_\delta \to k$ uniformly on the compact subinterval of (t_1, t_2), the lemma follows by the classical Kamke theorem for ordinary differential equations. □

Let us denote by $z_s(x, t_1, t_2, m, k)$ the function z_s defined previously depending on the parameters t_1, t_2, m, k. By Theorem B we have that the principal solutions of (10) are oscillatory. By Lemma 5 it follows that if η has the D-property, then the solution of CP (12), (12.1) is oscillatory. Hence, by Lemma 4, the solution of (13), (13.1) is oscillatory for every $k>0$. Therefore we can consider the function $h:\mathbf{R}^+\to\mathbf{R}^+$, where $h(k)$ is defined to be the first positive number such that the mth derivative of the solution of (13), (13.1) vanishes at the point $t_1+h(k)$.

LEMMA 7. *For $k\to 0^+$, $k^{-1/n}h(k)\to\alpha>0$. Moreover, for every i, $0\leqq i\leqq n-1$, $i\neq m-1$,*

$$\frac{\dfrac{\partial^{(i)}z_s}{\partial x^{(i)}}(t_1+h(k)+0, t_1, t_1+h(k), m, k)}{\dfrac{\partial^{(m-1)}z_s}{\partial x^{(m-1)}}(t_1+h(k)+0, t_1, t_1+h(k), m, k)}\to 0.$$

Proof. Without loss of generality we may suppose that $t_1=0$. Let $v=k^{-1/n}$ and let $u_i(x, v)$, $1\leqq i\leqq n$, be the principal solutions of $y^{(n)}+v^n q_0 y=0$ with initial point zero. Morever, let $\eta_{i+1}=z_s^{(i)}(0)$, $i=0,\cdots,n-1$. For $x_0\in(0, t(k))$ and for $j=0,\cdots,n-1$ we have

$$\begin{aligned}\frac{\partial^{(j)}z_s}{\partial x^{(j)}}(x_0, 0, h(k), m, k) &= \sum_{i=1}^{m}\eta_i\frac{\partial^{(j)}u_i}{\partial x^{(j)}}(x_0, v)+\sum_{i=m+1}^{n}\eta_i v^n\frac{\partial^{(j)}u_i}{\partial x^{(j)}}(x_0, v)\\ &= \sum_{i=1}^{m}\eta_i v^{1-i+j}\frac{\partial^{(j)}u_i}{\partial x^{(j)}}(x_0 v, 1)\\ &\quad+\sum_{i=m+1}^{n}\eta_i v^{n+1-i+j}\frac{\partial^{(j)}u_i}{\partial x^{(j)}}(x_0 v, 1).\end{aligned}$$

Therefore for $j\leqq m-1$ we have

$$\begin{aligned}\frac{\partial^{(j)}z_s}{\partial x^{(j)}}(t(k)+0, 0, h(k), m, k) &= \sum_{i=1}^{m}\eta_i v^{1-i+j}\frac{\partial^{(j)}u_i}{\partial x^{(j)}}(vh(k), 1)\\ &\quad+\sum_{i=m+1}^{n}\eta_i v^{n+1-i+j}\frac{\partial^{(j)}u_i}{\partial x^{(j)}}(vh(k), 1),\end{aligned}\tag{*}$$

while for $j\geqq m$ we have

$$\begin{aligned}\frac{\partial^{(j)}z_s}{\partial x^{(j)}}(t(k)+0, 0, h(k), m, k) &= \sum_{i=1}^{m}\eta_i v^{1-i+j-n}\frac{\partial^{(j)}u_i}{\partial x^{(j)}}(vh(k), 1)\\ &\quad+\sum_{i=m+1}^{n}\eta_i v^{1-i+j}\frac{\partial^{(j)}u_i}{\partial x^{(j)}}(vh(k), 1).\end{aligned}\tag{**}$$

By definition of $h(k)$, $vh(k)$ is the first positive zero of $(\partial^{(m)}z_s/\partial x^{(m)})(x/v, 0, h(k), m, k)$. Since for $x_0\in(0, vh(k))$ we have

$$\begin{aligned}\frac{\partial^{(m)}z_s}{\partial x^{(m)}}(x_0/v, 0, h(k), m, k) &= \sum_{i=1}^{m}\eta_i v^{1-i+m}\frac{\partial^{(m)}u_i}{\partial x^{(m)}}(x_0, 1)\\ &\quad+\sum_{i=m+1}^{n}\eta_i v^{n+1-i+m}\frac{\partial^{(m)}u_i}{\partial x^{(m)}}(x_0, 1)\end{aligned}$$

it follows that $vh(k)$ converges to the first positive zero of $(\partial^{(m)}u_t/\partial x^{(m)})(x, 1)$, where if $\eta_m \neq 0$ then $t = m+1$, while if $\eta_m = 0$ then $t = m+2$ (1 when $m = n-1$). The lemma now follows trivially by (*), (**). □

Proof of Theorem 5. Suppose that $k = n-1$, so that $q_0 > 0$.

Let s be the unique index such that $1 \leqq s \leqq n$ and $s-1$ does not belong to $\{i_1, \cdots, i_{n-1}\}$. Every solution of (10), (10.1) is a multiplication by a nonvanishing constant of the principal solutions u_s of (10) and initial point a. Let b_1 be the first zero of $u_s^{(j_1)}$ in the interval $(a, b]$.

We first consider the case $m = 0$. By Theorem 2 of [8] we have that there exists a function q such that $q(x) \geqq q_0$ on $[a, b_1]$ and such that the BVP

$$y^{(n)} + q(x)y = 0, \tag{14}$$

$$y^{(i_1)}(a) = \cdots = y^{(i_{n-1})}(a) = 0, \qquad y^{(j_1)}(c) = 0, \tag{14.1}$$

has no solution for every $c \in (a, b_1]$. Moreover, an examination of the proof of the theorem shows that q can be chosen such that $q(x) = q_0$ in a neighborhood of the point a and of the point b_1. Let $\tilde{p}_0 = q_0/q$. The BVP (14), (14.1) is equivalent to (11), (11.1), where $m = 0$; hence the function $\tilde{p}_0$ satisfies the theorem.

Suppose now that $m \geqq 1$. Since system (10.1) is not m-admissible, by Lemma 2, $u_s^{(m)}$ vanishes on (a, b_1). Moreover, on this interval $u_s^{(m)}$ has only one zero x_1, otherwise by Lemma 4, $u_s^{(j_1)}$ vanishes on (a, b_1) in contradiction to the definition of b_1. We denote with v_ξ the solution of CP (12), (12.1), where $q = q_0$ and $\eta = e_m$. By Lemma 4 the vector $\eta = (u_s(x_1), \cdots, u_s^{(n-1)}(x_1))$ has the strict D-property and $r(\eta) = m$. By Lemma 5 the first zero of $v_{x_1}^{(j_1)}$ on $(x_1, +\infty)$ is greater than b_1. Hence there exists a point $x_2 < x_1$ and a number $\delta > 0$ such that for every $\xi \in [x_2, x_1]$ the first zero of $v_\xi^{(j_1)}$ is greater than $b_1 + \delta$. Let us now consider the function $z_s(x, x_2, x_2 + h(k), m, k)$ defined previously. By Lemma 7 for $k \to 0^+$ we have $h(k) \to 0$ and moreover,

$$\frac{\dfrac{\partial^{(i)} z_s}{\partial x^{(i)}}(h(k)+0, x_2, x_2+h(k), m, k)}{\dfrac{\partial^{(m-1)} z_s}{\partial x^{(m-1)}}(h(k)+0, x_2, x_2+h(k), m, k)} \to 0 \quad \text{for every } i \neq m-1.$$

By the theorem on the continuous dependence on initial conditions and by definition of x_2, we have that there exists $k_0 < 1$ such that $x_2 + h(k_0) < x_1$ and such that the first zero of $\partial^{(j_1)} z_s/\partial x^{(j_1)}(x, x_2, x_2 + h(k_0), m, k_0)$ is greater than $b_1 + \delta/2$. Let ε_0 be a positive number such that for every $x \in [x_2, b_1]$, $\varepsilon_0 < k_0 |z_s^{(j_1)}(x, x_2, x_2 + h(k_0), m, k_0)|$. In correspondence to ε_0 and $z_s(x, x_2, x_2 + h(k_0), m, k_0)$ there exists a function r satisfying Lemma 6. We put $\tilde{p}_m = r$. By the definition of the function z_s and by Lemma 6, we have that the j_1th derivative of the principal solution w_s of (11) does not vanish on $(a, b_1]$, and hence the point b_1 and the function $\tilde{p}_m$ satisfy the theorem.

Suppose now that $k = 1$.

As remarked in [8] this case can be trivially reduced to the previous one considering the equivalence of BVP (11), (11.1) to the BVP

$$(\tilde{p}_m(a + c - x)z^{(m)})^{(n-m)} + (-1)^n q_0 z = 0,$$

$$M_j z(a) = 0, \quad j \in \{j_1, \cdots, j_{n-k}\}, \qquad M_i z(c) = 0, \quad i \in \{i_1, \cdots, i_k\}. \qquad \square$$

Proof of Theorem 6. Let λ_0 and μ_0 be, respectively, the first positive eigenvalue of EP (10λ), (10.1) and of EP (11λ), (11.1). Moreover, let b_1 and $\tilde{p}_m$ be the point and

the function arising from the application of Theorem 5 to the BVP given by

$$y^{(n)}+\lambda_0 q_0 y=0 \tag{15}$$

and boundary conditions (10.1).

By Theorem B(d), λ_0 is a decreasing function of the boundary points. Therefore b_1 cannot belong to (a, b), so we have $b_1=b$. Hence the BVP given by

$$(\tilde{p}_m u^{(m)})^{(n-m)}+\lambda_0 q_0 u=0$$

and system (11.1) has no solution for every $c\in(a, b]$. Again by Theorem B(d), we have that the first positive eigenvalue of EP given by

$$(\tilde{p}_m u^{(m)})^{(n-m)}+\mu\lambda_0 q_0 u=0 \tag{16}$$

and system (11.1) is greater than 1. Since the first positive eigenvalue of (16), (11.1) is μ_0/λ_0, we have $\mu_0>\lambda_0$. □

Finally, we consider an example showing that if we drop hypothesis (Q_1) in Theorem 1, then BVP (4), (4.1) can have no solution for every $c\in(a, b)$, even if $\tilde{p}\not\equiv p$.

Let λ_0 be the first eigenvalue of:

$$y'''+\lambda|x|y=0, \qquad y(a)=0, \quad y''(a)=0, \quad y'(b)=0,$$

where $a<0$ and $b>0$, and let q be the function defined in the following way:

$$q(x)=\begin{cases}|x|\lambda_0 & \text{if } x\leqq 0,\\ 0 & \text{if } x>0.\end{cases}$$

If y_0 is an eigenfunction corresponding to λ_0 then, by Lemma 4, we have $y_0'(0)y_0''(0)<0$.

Let us consider the BVP

$$(p(x)u')''+q(x)u=0, \tag{17}$$

$$u(a)=0, \quad (pu')'(a)=0, \quad pu'(c)=0, \tag{17.1}$$

where $c>0$ and p is a function of class C^∞ such that $p(x)=1$ for $x\leqq 0$ and $p(x)>0$ for $x>0$.

If u_0 is a solution of (17), (17.1) then there exists $k\neq 0$ such that $u_0(x)=ky_0(x)$ for $a\leqq x\leqq 0$. Therefore we have

$$p(x)u'(x)=\begin{cases}ky_0'(x) & \text{for } x\leqq 0,\\ k(y_0''(0)x+y_0'(0)) & \text{for } x>0.\end{cases}$$

By definition of y_0 we have that y_0' does not vanish on $[a, 0]$. Therefore, for every function p, pu' vanishes only at the point $b_0=-y_0'(0)/y_0''(0)>0$. Hence, if we consider two such functions p_1, p_2 such that $0<p_2(x)<p_1(x)$ for $x>0$, then setting $p=p_1$, BVP (17), (17.1) has a solution for $c=b_0$, whereas setting $p=p_2$ BVP (17), (17.1) has no solution for every $c<b_0$.

REFERENCES

[1] S. AHMAD AND A. C. LAZER, *On n-th order Sturmian theory*, J. Differential Equations, 35 (1980), pp. 87-112.

[2] ———, *On an extension of Sturm comparison theorem*, SIAM J. Math. Anal., 12 (1981), pp. 1-9.

[3] G. J. BUTLER AND L. H. ERBE, *Integral comparison theorems and extremal points for linear differential equations*, J. Differential Equations, 47 (1983), pp. 214-226.

[4] W. A. COPPEL, *Disconjugacy*, Lecture Notes in Math. 20, Springer-Verlag, Berlin, New York, 1955.

[5] U. ELIAS, *Eigenvalue problems for the equation* $Ly+\lambda p(x)y=0$, J. Differential Equations, 29 (1978), pp. 28-57.

[6] U. ELIAS, *Oscillatory solutions and extremal points for linear differential equations*, Arch. Rational Math. Soc., 76 (1979), pp. 51-59.

[7] ———, *The extremal solutions of the equation* $Ly+p(x)y=0$, II, J. Math. Anal. Appl., 55 (1976), pp. 253-265.

[8] M. GAUDENZI, *On an eigenvalue problem of Ahmad-Lazer for ordinary differential equations*, Proc. Amer. Math. Soc., 99 (1987), pp. 237-243.

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

[10] Z. NEHARI, *Disconjugate linear differential operators*, Trans. Amer. Math. Soc., 129 (1969), pp. 500-516.

[11] K. SCHMITT, *Boundary value problems and comparison theorems for ordinary differential equations*, SIAM J. Appl. Math., 26 (1979), pp. 500-516.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 995-1018, July 1990
© 1990 Society for Industrial and Applied Mathematics
010

BESSEL FUNCTIONS OF PURELY IMAGINARY ORDER, WITH AN APPLICATION TO SECOND-ORDER LINEAR DIFFERENTIAL EQUATIONS HAVING A LARGE PARAMETER*

T. M. DUNSTER†

Abstract. Bessel functions of purely imaginary order are examined. Solutions of both the modified and unmodified Bessel equations are defined which, when their order is purely imaginary and their argument is real and positive, are pairs of real numerically satisfactory functions. Recurrence relations, analytic continuation formulas, power series representations, Wronskian relations, integral representations, behavior at singularities, and asymptotic forms of the zeros are derived for these numerically satisfactory functions. Also, asymptotic expansions in terms of elementary and Airy functions are derived for the Bessel functions when their order is purely imaginary and of large absolute value.

Second-order linear ordinary differential equations having a large parameter and a simple pole are then examined, for the case where the exponent of the pole is complex. Asymptotic expansions are derived for the solutions in terms of the numerically satisfactory Bessel functions of purely imaginary order.

Key words. asymptotic analysis, Bessel functions, ordinary differential equations, zeros

AMS(MOS) subject classifications. primary 33A40; secondary 34E20

1. Introduction and summary. The purpose of this paper is to investigate solutions of both the unmodified and the modified Bessel equations (see (2.1) and (3.1)). We consider the case where the parameter μ in the equations is purely imaginary, so that the solutions are of purely imaginary order.

Consider first the asymptotic behavior of Bessel functions. This is an area that has been extensively studied, reflecting the importance of Bessel functions in many areas of mathematics and physics. Uniform asymptotic expansions of modified and unmodified Bessel functions of complex argument and large positive order are available in terms of both elementary and Airy functions, see § 7 of Chap. 10 and § 10 of Chap. 11 in Olver (1974). (We will refer to Olver's book frequently, and therefore here and throughout we use the abbreviation "Chap." to refer to a chapter of that text.) Expansions for complex orders with positive (nonzero) real part are also available: see § 8 of Chap. 10.

When the parameter $\mu \equiv i\nu$ is purely imaginary, however, the picture is less complete; only the modified Bessel function of the third kind $K_{i\nu}(z)$ seems to have been extensively studied. Uniform asymptotic expansions in terms of Airy functions have been derived for $K_{i\nu}(\nu z)$, ν real and large, by Balogh (1967). The positive zeros of $K_{i\nu}(z)$ have also been investigated by a number of authors (see, e.g., Ferreira and Sesma (1970), Laforgia (1986)). For other asymptotic results concerning Bessel functions of purely imaginary order, see Jeffreys (1962, pp. 90-91) and Falcão (1973).

Regarded as a Bessel function of purely imaginary order, the function $K_{i\nu}(x)$ is unique in two respects. First, it alone of the standard Bessel functions is real when the argument x is positive: the Bessel functions $J_{i\nu}(x)$, $Y_{i\nu}(x)$, $H^{(1)}_{i\nu}(x)$, $H^{(2)}_{i\nu}(x)$, and the modified Bessel function $I_{i\nu}(x)$ are all complex when ν and x are real and nonzero.

Second, $K_{i\nu}(x)$ is recessive as $x \to +\infty$, a property that makes the function useful in certain physical problems, such as hydrodynamical, quantum mechanical, and

* Received by the editors May 2, 1988; accepted for publication (in revised form) August 24, 1989. This research was partly supported by the University of British Columbia, Vancouver, British Columbia, Canada.

† Department of Mathematical Sciences, San Diego State University, San Diego, California 92182.

diffraction theories. Also, this property allows us to readily identify the function with asymptotic solutions of (2.1) that it satisfies.

An example in which Bessel functions of purely imaginary order play an important role in quantum mechanics is the problem of s-wave scattering by a reduced exponential potential; see Kogan and Galitsky (1963, pp. 334–341) and Joachain (1975, p. 317). For other applications see Hemker (1974).

On account of the two above-mentioned properties, $K_{i\nu}(x)$ remains one of a pair of solutions of (3.1) on the x interval $(0,\infty)$ that are numerically satisfactory; see Miller (1950). $K_{i\nu}(x)$ is oscillatory in a neighborhood of the $x=0$, and is exponential in a neighborhood of $x=\infty$. An appropriate numerically satisfactory companion would be a real solution which is of equal amplitude and $\pi/2$ out of phase in the oscillatory region. We introduce a function, denoted by $L_{i\nu}(x)$, which fulfills these criteria (see (2.2)).

Solutions of the unmodified Bessel equation (3.1) that are real when μ is purely imaginary, and $z\equiv x$ is positive, are oscillatory throughout $0<x<\infty$. We introduce two real solutions, denoted by $F_{i\nu}(x)$ and $G_{i\nu}(x)$, that are $\pi/2$ out of phase on $0<x<\infty$, have equal amplitudes of oscillation at $x=\infty$ for all $\nu>0$, and have asymptotically equal amplitudes of oscillation throughout $0<x<\infty$ as $\nu\to\infty$ (see (3.2) and (3.3)).

The plan of this paper is as follows. In §§ 2 and 3 we derive a number of results concerning $K_\mu(z)$, $L_\mu(z)$, $F_\mu(z)$, and $G_\mu(z)$, most of which pertain to μ being purely imaginary. We record recurrence relations, analytic continuation formulas, power series representations, Wronskian relations, connection formulas, integral representations, and asymptotic behavior at $z=0$ and ∞. These results follow from standard results concerning Bessel functions, the latter being found, for example, by perusing Olver (1974), and in most instances details of their derivations have been omitted.

In §§ 2 and 3 the zeros of the four functions are also examined; asymptotic formulas are derived for the zeros including those of $K_{i\nu}(x)$, which are of importance in certain physical problems, such as in quantum mechanics (see Gray, Mathews, and MacRobert (1952)).

In §§ 4 and 5 we examine the asymptotic behavior of the four functions as $\nu\to\infty$. As has already been noted, the modified Bessel function $K_{i\nu}(\nu z)$ has been studied by Balogh (1967). The corresponding asymptotic expansion for $L_{i\nu}(\nu z)$ (as well as that for $I_{i\nu}(\nu z)$) is derived, using the theory of Chap. 11. The theory of Chap. 10 is applied to deriving asymptotic expansions, involving elementary functions, for $F_{i\nu}(\nu z)$ and $G_{i\nu}(\nu z)$ (as well as for the Hankel functions $H_{i\nu}^{(1)}(\nu z)$ and $H_{i\nu}^{(2)}(\nu z)$).

One example of a useful application for these asymptotic results is to problems of high-frequency wave propagation in inhomogeneous media with linear velocity profiles (see, for example, Gupta (1965)). For detailed discussions of this class of problem see Brekhovskikh (1960). In his paper Gupta uses expansions (29) for $\arg\nu=\pi/2$, and although this is not justified, we shall see that the first of these represents the (dominant) real part when $y=-i\sigma$, $1<\sigma<\infty$.

In § 7 asymptotic solutions are constructed for equations of the form

$$\frac{d^2w}{dx^2}=\{u^2f(x)+g(x)\}w, \tag{1.1}$$

in which u is a large parameter, the independent variable x lies in an open (finite or infinite) interval, and at some point $x=x_0$, $f(x)$ has a simple pole and $(x-x_0)^2g(x)$ is analytic. It is supposed that there are no other transition points (zeros of $f(x)$ or singularities) in the x interval under consideration.

We consider the case where

$$\lim_{x\to x_0}(x-x_0)^2 g(x) \equiv -\frac{\nu^2+1}{4} < -\frac{1}{4} \qquad (\nu>0), \tag{1.2}$$

which corresponds to the exponents of the pole $x = x_0$ being complex (see, e.g., § 4 of Chap. 5).

The complementary problem, where the exponents are real, has been tackled by Olver (see §§ 1-4 of Chap. 12). We proceed in a manner similar to Olver's. By means of the Liouville transformation, formulas (2.02) and (2.03) of Chap. 12, our equation (1.1) is transformed to the form (7.1), where ν is positive (compare (2.05) of Chap. 12). Equation (7.1) is the focus of our attention; asymptotic solutions are constructed in terms of the Bessel functions of purely imaginary order $K_{i\nu}(x)$, $L_{i\nu}(x)$, $F_{i\nu}(x)$, and $G_{i\nu}(x)$. Using auxiliary functions for these four functions (given in § 6), we derive error bounds for the asymptotic expansions.

It will be assumed that the reader is familiar with the results in Chaps. 10 and 11 and §§ 1-5 of Chap. 12.

2. Modified Bessel functions of purely imaginary order. The modified Bessel functions $I_\mu(z)$ and $K_\mu(z)$ compose a numerically satisfactory pair of solutions of the modified Bessel equation

$$\frac{d^2w}{dz^2}+\frac{1}{z}\frac{dw}{dz}-\left(1+\frac{\mu^2}{z^2}\right)w=0 \tag{2.1}$$

in the half-plane $|\arg z| \leqq \pi/2$, for all complex values of μ such that $\operatorname{Re}\mu \geqq 0$. By "numerically satisfactory" we mean a pair of linearly independent solutions that satisfy the criteria of Miller (1950). When μ is purely imaginary, however, the function $I_\mu(z)$ has the undesirable property of being complex on the positive real z axis. Therefore we now introduce the following function:

$$L_\mu(z)=\frac{\pi i}{2\sin(\mu\pi)}\{I_\mu(z)+I_{-\mu}(z)\} \qquad (\mu\neq 0), \tag{2.2}$$

which will be seen to be an appropriate numerically satisfactory companion to $K_{i\nu}(x)$, where ν is real and nonzero, and x is real and positive. Note that $L_\mu(z)$ is not defined when $\mu = 0$. It is not possible to define a numerically satisfactory companion to $K_{i\nu}(x)$ which remains finite as $\nu \to 0$.

The definition of $L_\mu(z)$ should be compared with the definition of $K_\mu(z)$:

$$K_\mu(z)=\frac{\pi}{2\sin(\mu\pi)}\{I_{-\mu}(z)-I_\mu(z)\}. \tag{2.3}$$

The purpose of this section is to record some relevant properties of $L_\mu(z)$ and $K_\mu(z)$, with emphasis on the case where μ is purely imaginary. Throughout, μ denotes a complex parameter, and ν denotes *a positive (nonzero) parameter.* When the independent variable z is real and positive we denote it by x.

Recurrence relations. The functions $e^{\mu\pi i}K_\mu(z)$ and $e^{\mu\pi i}L_\mu(z)$ satisfy the same recurrence relations as $I_\mu(z)$, viz.

$$\begin{aligned} &I_{\mu-1}(z)-I_{\mu+1}(z)=(2\mu/z)I_\mu(z), \qquad I_{\mu-1}(z)+I_{\mu+1}(z)=2I'_\mu(z),\\ &I_{\mu+1}(z)=-(\mu/z)I_\mu(z)+I'_\mu(z), \qquad I_{\mu-1}(z)=(\mu/z)I_\mu(z)+I'_\mu(z). \end{aligned} \tag{2.4}$$

Analytic continuation. For any integer m

$$K_\mu(z\,e^{m\pi i}) = \cos(m\mu\pi)K_\mu(z) - \sin(m\mu\pi)L_\mu(z), \tag{2.5a}$$

$$L_\mu(z\,e^{m\pi i}) = \cos(m\mu\pi)L_\mu(z) + \sin(m\mu\pi)K_\mu(z). \tag{2.5b}$$

Power series representation. A power series expansion for $L_\mu(z)$ in ascending powers of z is readily derived from (2.2) together with the well-known power series for $I_{\pm\mu}(z)$ (see formula (10.01) of Chap. 2). When $\mu = i\nu$ and $z = x$ this power series can be expressed as

$$L_{i\nu}(x) = \left(\frac{\nu\pi}{\sinh(\nu\pi)}\right)^{1/2} \sum_{s=0}^{\infty} \frac{(x^2/4)^s \cos(\nu \ln(x/2) - \phi_{\nu,s})}{s![(\nu^2)(1^2+\nu^2)\cdots(s^2+\nu^2)]^{1/2}}, \tag{2.6}$$

where

$$\phi_{\nu,s} = \arg\{\Gamma(1+s+i\nu)\}. \tag{2.7}$$

(For each s we define the branch of (2.7) so that $\phi_{\nu,s}$ is continuous for $0 < \nu < \infty$, with $\lim_{\nu\to 0} \phi_{\nu,s} = 0$.)

From the definition (2.3) of $K_\mu(z)$ we derive in a similar manner

$$K_{i\nu}(x) = -\left(\frac{\nu\pi}{\sinh(\nu\pi)}\right)^{1/2} \sum_{s=0}^{\infty} \frac{(x^2/4)^2 \sin(\nu \ln(x/2) - \phi_{\nu,s})}{s![(\nu^2)(1^2+\nu^2)\cdots(s^2+\nu^2)]^{1/2}}. \tag{2.8}$$

Connection formulas.

$$L_{-\mu}(z) = -L_\mu(z), \qquad K_{-\mu}(z) = K_\mu(z). \tag{2.9}$$

Wronskian.

$$\mathcal{W}\{K_\mu(z), L_\mu(z)\} = \frac{\pi i}{\sin(\mu\pi)z}. \tag{2.10}$$

Integral representations. To derive an integral representation for $L_{i\nu}(z)$ we first express the function as a linear combination of Hankel functions. From the definition of $L_{i\nu}(z)$ and § 8.1 of Chap. 7 we obtain

$$L_{i\nu}(z) = \frac{\pi}{2\sinh(\nu\pi)}\{e^{-(\nu\pi/2)}\cosh(\nu\pi)H_{i\nu}^{(1)}(z\,e^{i\pi/2}) + e^{(\nu\pi/2)}H_{i\nu}^{(2)}(z\,e^{i\pi/2})\}. \tag{2.11}$$

Next, the Hankel functions in (2.11) are expressed by their Sommerfeld integral representations (equation (4.19) of Chap. 7, with $\alpha = \pi/2$). The resulting integral representation for $L_{i\nu}(z)$ can be re-expressed, via a splitting into three integrals followed by appropriate changes of integration variables, in the following form:

$$\begin{aligned} L_{i\nu}(z) = [\sinh(\nu\pi)]^{-1} &\int_0^{\pi} e^{z\cos\theta}\cosh(\nu\theta)\,d\theta \\ &- \int_0^{\infty} e^{-z\cosh t}\sin(\nu t)\,dt, \qquad |\arg z| < \pi/2. \end{aligned} \tag{2.12}$$

It is immediately seen from (2.12) that $L_{i\nu}(x)$ is real for $0 < x < \infty$. The modified Bessel function $K_{i\nu}(z)$ has the known integral representation

$$K_{i\nu}(z) = \int_0^{\infty} e^{-z\cosh t}\cos(\nu t)\,dt, \qquad |\arg z| < \pi/2, \tag{2.13}$$

and from this it is seen that $K_{i\nu}(x)$, too, is real for $0 < x < \infty$.

Behavior at the singularities $z=0, \infty$. If $v(>0)$ is fixed and $x \to 0^+$, then

$$K_{iv}(x) = -\left(\frac{\pi}{v \sinh(v\pi)}\right)^{1/2} \{\sin(v \ln(x/2) - \phi_{v,0}) + O(x^2)\}, \tag{2.14}$$

$$L_{iv}(x) = \left(\frac{\pi}{v \sinh(v\pi)}\right)^{1/2} \{\cos(v \ln(x/2) - \phi_{v,0}) + O(x^2)\}. \tag{2.15}$$

(Note that the amplitudes of oscillation of $L_{iv}(x)$ and $K_{iv}(x)$ in a neighborhood of the origin become unbounded as $v \to 0$.) As $z \to \infty$

$$K_{iv}(z) = \left(\frac{\pi}{2z}\right)^{1/2} e^{-z}\left\{1 + O\left(\frac{1}{z}\right)\right\}, \qquad |\arg z| \leqq 3\pi/2 - \delta, \tag{2.16}$$

$$L_{iv}(z) = \frac{1}{\sinh(v\pi)}\left(\frac{\pi}{2z}\right)^{1/2} e^{z}\left\{1 + O\left(\frac{1}{z}\right)\right\}, \qquad |\arg z| \leqq \pi/2 - \delta, \tag{2.17}$$

where δ is *an arbitrary small positive constant* (a convention used throughout). It should be emphasized that in (2.17) we have neglected exponentially small contributions (in Poincaré's sense), and as such the error term in this asymptotic formula can be large near the boundary of the region of validity. Inclusion of the exponentially small terms will result both in an extension of the region of validity, and increased accuracy (cf. Exercise 13.2 of Chap. 7).

Zeros. When $v > 0$ it is known that $K_{iv}(x)$ has an infinite number of simple positive zeros in $0 < x < v$, and no zeros in $v \leqq x < \infty$. We denote the zeros by $\{k_{v,s}\}_{s=1}^{\infty}$, such that

$$v > k_{v,1} > k_{v,2} > k_{v,3} > \cdots > 0, \tag{2.18}$$

$$\lim_{s\to\infty} k_{v,s} = 0. \tag{2.19}$$

LEMMA 1. *When* $v > 0$, $L_{iv}(x)$ *has an infinite number of simple positive zeros, denoted by* $\{l_{v,s}\}_{s=1}^{\infty}$, *say, such that*

$$l_{v,1} > k_{v,1} > l_{v,2} > k_{v,2} > l_{v,3} > \cdots > 0, \tag{2.20}$$

$$\lim_{s\to\infty} l_{v,s} = 0. \tag{2.21}$$

Proof. The asymptotic behavior of $L_{iv}(x)$ near $x = 0$ (see (2.15)) shows that the function has an infinite number of positive zeros. Using the Wronskian relation (2.10) and arguing along the lines of the proof of Theorem 7.1 of Chap. 7 we see that the zeros $k_{v,s}$, $l_{v,s}$ $(s = 0, 1, 2, \cdots)$ are interlaced. It remains then to show that $l_{v,1} > k_{v,1}$. Suppose that $k_{v,1} > l_{v,1} (> k_{v,2})$. From (2.16) it is seen that $K_{iv}(x)$ is positive for $x > k_{v,1}$, and therefore negative in $(k_{v,2}, k_{v,1})$; in particular $K_{iv}(l_{v,1}) < 0$. From (2.17) it is seen that $L_{iv}(x)$ is positive in $(l_{v,1}, \infty)$, and therefore $L'_{iv}(l_{v,1}) > 0$. Thus the assumption $k_{v,1} > l_{v,1}$ implies

$$K_{iv}(l_{v,1}) L'_{iv}(l_{v,1}) < 0,$$

which contradicts the fact that the Wronskian (2.10) is positive for $0 < x < \infty$. Thus $l_{v,1} > k_{v,1}$ as asserted. □

Asymptotic approximations for the zeros $\{k_{v,s}\}_{s=1}^{\infty}$ of $K_{iv}(x)$ can be derived from (2.14), and also from the asymptotic expansions given in § 4 (see (4.7) and (4.8)). We now record asymptotic approximations for the zeros which can be derived from these results. First, consider the asymptotic behavior of the zeros as $v \to \infty$: from (4.3), (4.7), and (4.8), together with the theory of § 6 of Chap. 11, we can show that

$$k_{v,s} = vX(-v^{-2/3}a_s) + s^{-1/3}O(v^{-2/3}) + O(v^{-1}), \tag{2.22}$$

as $\nu \to \infty$, uniformly for all positive integers s. Here $X(\zeta)$ is defined implicitly by the equation

$$\frac{2}{3}\zeta^{3/2} = \ln\left\{\frac{1+(1-X^2)^{1/2}}{X}\right\} - (1-X^2)^{1/2}, \tag{2.23}$$

and $\{a_s\}_{s=1}^{\infty}$ denote the (negative) zeros of the Airy function $\mathrm{Ai}(x)$, with the usual convention

$$0 > a_1 > a_2 > \cdots. \tag{2.24}$$

For *fixed* s the sth zero of $K_{i\nu}(x)$ takes the simplified form

$$k_{\nu,s} = \nu + a_s(\nu/2)^{1/3} + \tfrac{3}{20}a_s^2(\nu/2)^{-1/3} + O(\nu^{-2/3}), \tag{2.25}$$

as $\nu \to \infty$.

Next, we consider the form of the zeros for fixed ν, as $s \to \infty$. We know that $k_{\nu,s} \to 0$ in this case; from the first two terms in the power series (2.8) we find that as $s \to \infty$ $(x \to 0)$

$$k_{\nu,s} = 2\, e^{-(1/\nu)(s\pi - \phi_{\nu,0})}\left\{1 + \frac{e^{-(2/\nu)(s\pi-\phi_{\nu,0})}}{(1+\nu^2)} + O(e^{-(4s\pi/\nu)})\right\}, \tag{2.26}$$

for fixed ν. We remark that it is not obvious that the right-hand side (RHS) of (2.26) is the sth zero of $K_{i\nu}(x)$, as opposed to, say, the $(s+1)$th zero. We now show that the RHS of (2.26) indeed represents $k_{\nu,s}$. To do so consider (2.22): this is uniformly valid for all integers s, and therefore we can consider the limiting form of this expression as $s \to \infty$, with ν large but now assumed fixed. On employing the approximation

$$a_s = -(3\pi(4s-1)/8)^{2/3} + O(s^{-4/3}) \tag{2.27}$$

(see (5.05) of Chap. 11) together with (2.23) we find that

$$\nu X(-\nu^{-2/3}a_s) \sim 2\, e^{-(1/\nu)(s\pi - \nu\ln\nu + \nu - \pi/4)} \quad \text{as } s \to \infty. \tag{2.28}$$

From the definition (2.7) of $\phi_{\nu,s}$, and the asymptotic form (see Abramowitz and Stegun (1965, p. 257))

$$\arg\{\Gamma(iy)\} \sim y\ln(y) - y - \frac{\pi}{4} \qquad (y \to +\infty)$$

we find that for large ν

$$\phi_{\nu,0} \sim \nu\ln(\nu) - \nu + \frac{\pi}{4}. \tag{2.29}$$

Thus on comparing (2.22), (2.28), and (2.29) with the RHS of (2.26) we deduce that the latter represents $k_{\nu,s}$ for large fixed ν and $s \to \infty$. By a continuity argument this is true for nonlarge values of ν as well.

Finally, we record corresponding asymptotic forms for $\{l_{\nu,s}\}_{s=1}^{\infty}$. As $\nu \to \infty$

$$l_{\nu,s} = \nu X(-\nu^{-2/3}b_s) + s^{-1/3}O(\nu^{-2/3}) + O(\nu^{-1}), \tag{2.30}$$

uniformly for all positive integers s. Here $X(\zeta)$ is again given by (2.23), and $\{b_s\}_{s=1}^{\infty}$ denote the (negative) zeros of the Airy function $\mathrm{Bi}(x)$, in ascending order of absolute magnitude.

For fixed s and $\nu \to \infty$

$$l_{\nu,s} = \nu + b_s(\nu/2)^{1/3} + \tfrac{3}{20}b_s^2(\nu/2)^{-1/3} + O(\nu^{-2/3}). \tag{2.31}$$

For fixed ν and $s \to \infty$

$$l_{\nu,s} = 2\, e^{-(1/\nu)((s-1/2)\pi - \phi_{\nu,0})}\left\{1 + \frac{e^{-(2/\nu)((s-1/2)\pi-\phi_{\nu,0})}}{(1+\nu^2)} + O(e^{-(4\pi/\nu)(s-1/2)})\right\}. \tag{2.32}$$

3. Unmodified Bessel functions of purely imaginary order. Standard solutions of the unmodified Bessel equation

$$\frac{d^2w}{dz^2}+\frac{1}{z}\frac{dw}{dz}+\left(1-\frac{\mu^2}{z^2}\right)w=0, \tag{3.1}$$

are the Bessel functions of the first and second kinds $J_\mu(z)$, $Y_\mu(z)$, and the Bessel functions of the third kind (Hankel functions) $H_\mu^{(1)}(z)$, $H_\mu^{(2)}(z)$. The characterizing properties of these functions are the following:

(i) $J_\mu(z)$ is recessive at the regular singularity $z=0$ when $\operatorname{Re}\mu>0$ or $\mu=0$, and moreover is real on the positive real z axis when μ is real.

(ii) $Y_\mu(z)$ is real for positive z and real μ, and for large positive z has the same amplitude of oscillation as $J_\mu(z)$ and is out of phase by $\pi/2$.

(iii) For all μ, $H_\mu^{(1)}(z)$ is recessive at infinity in the sector $\delta\leqq\arg z\leqq\pi-\delta$, and $H_\mu^{(2)}(z)$ is recessive in the conjugate sector.

Thus, when μ is real and nonnegative, $J_\mu(z)$ and $Y_\mu(z)$ are a numerically satisfactory pair on $0<x<\infty$, and when $\operatorname{Re}\mu>0$ or $\mu=0$, $J_\mu(z)$ and $H_\mu^{(1)}(z)$ are a numerically satisfactory pair throughout the sector $0\leqq\arg z\leqq\pi$, $J_\mu(z)$, $H_\mu^{(2)}(z)$ being the numerically satisfactory pair in the conjugate sector.

When $\arg\mu=\pm\pi/2$ no solution is recessive at the origin, and the Hankel functions $H_\mu^{(1)}(z)$ and $H_\mu^{(2)}(z)$ compose a numerically satisfactory pair throughout $|\arg z|\leqq\pi$. However, these functions, as well as $J_\mu(z)$ and $Y_\mu(z)$, are not real on the real z axis when their orders are purely imaginary. We therefore now introduce two new Bessel functions that *are* real when $z\equiv x$ is positive and $\arg\mu=\pm\pi/2$, and moreover are numerically satisfactory when x and $|\mu|$ are not both small. We define

$$F_\mu(z)=\tfrac{1}{2}\{e^{\mu\pi i/2}H_\mu^{(1)}(z)+e^{-\mu\pi i/2}H_\mu^{(2)}(z)\}, \tag{3.2}$$

$$G_\mu(z)=\frac{1}{2i}\{e^{\mu\pi i/2}H_\mu^{(1)}(z)-e^{-\mu\pi i/2}H_\mu^{(2)}(z)\}. \tag{3.3}$$

From the relations $H_\mu^{(1)}(z)=\overline{H_{\bar\mu}^{(2)}(\bar z)}$, $H_{-\mu}^{(2)}(z)=e^{-\mu\pi i}H_\mu^{(2)}(z)$, where bars denote complex conjugates, it is readily verified that $F_{i\nu}(x)$ and $G_{i\nu}(x)$ are real for $x>0$. (Note that $F_0(z)=J_0(z)$ and $G_0(z)=Y_0(z)$.) Also, from the following alternative representations, which can be derived from standard results (see, e.g., (4.12), of Chap. 7)

$$F_\mu(z)=\tfrac{1}{2}\sec(\mu\pi/2)\{J_\mu(z)+J_{-\mu}(z)\}, \tag{3.4a}$$

$$G_\mu(z)=\tfrac{1}{2}\operatorname{cosec}(\mu\pi/2)\{J_\mu(z)-J_{-\mu}(z)\}, \tag{3.4b}$$

it is seen that $\cos(\mu\pi/2)F_\mu(z)$ and $\sin(\mu\pi/2)G_\mu(z)$ satisfy the same recurrence relations as $J_\mu(z)$, namely (2.3) with $I_{\mu+1}(z)$, $I_{\mu-1}(z)$, $I_\mu(z)$, $I'_\mu(z)$ replaced by $-J_{\mu+1}(z)$, $J_{\mu-1}(z)$, $J_\mu(z)$, $J'_\mu(z)$, respectively.

We now record other properties of $F_\mu(z)$ and $G_\mu(z)$.

Analytic continuation. For any integer m

$$F_\mu(z\,e^{m\pi i})=\cos(m\mu\pi)F_\mu(z)+i\sin(m\mu\pi)\tan(\mu\pi/2)G_\mu(z), \tag{3.5}$$

$$G_\mu(z\,e^{m\pi i})=i\sin(m\mu\pi)\cot(\mu\pi/2)F_\mu(z)+\cos(m\mu\pi)G_\mu(z). \tag{3.6}$$

Connection formulas.

$$F_{-\mu}(z)=F_\mu(z),\qquad G_{-\mu}(z)=G_\mu(z). \tag{3.7}$$

Power series representations. For purely imaginary order and positive argument we have

$$F_{i\nu}(x)=\left(\frac{2\nu\tanh(\nu\pi/2)}{\pi}\right)^{1/2} \cdot\sum_{s=0}^{\infty}\frac{(-1)^s(x^2/4)^s\cos(\nu\ln(x/2)-\phi_{\nu,s})}{s![(\nu^2)(1+\nu^2)\cdots(s^2+\nu^2)]^{1/2}}, \tag{3.8}$$

$$G_{i\nu}(x)=\left(\frac{2\nu\coth(\nu\pi/2)}{\pi}\right)^{1/2} \cdot\sum_{s=0}^{\infty}\frac{(-1)^s(x^2/4)^s\sin(\nu\ln(x/2)-\phi_{\nu,s})}{s![(\nu^2)(1+\nu^2)\cdots(s^2+\nu^2)]^{1/2}}, \tag{3.9}$$

where $\phi_{\nu,s}$ is defined by (2.7).

Wronskian.

$$\mathcal{W}\{F_\mu(z),G_\mu(z)\}=2/(\pi z). \tag{3.10}$$

Integral representations. The Schläfli-type representations are readily shown to be

$$F_\mu(z)=\frac{1}{2\pi i\cos(\mu\pi/2)}\int_{\infty-\pi i}^{\infty+\pi i}e^{z\sinh t}\cosh(\mu t)\,dt, \qquad |\arg z|<\pi/2, \tag{3.11}$$

$$G_\mu(z)=\frac{-1}{2\pi i\sin(\mu\pi/2)}\int_{\infty-\pi i}^{\infty+\pi i}e^{z\sinh t}\sinh(\mu t)\,dt, \qquad |\arg z|<\pi/2, \tag{3.12}$$

where the path of integration is indicated in Fig. 1.

For purely imaginary order these integrals can be re-expressed as

$$F_{i\nu}(z)=\frac{1}{\pi}\operatorname{sech}(\nu\pi/2)\int_0^{\pi}\cos(z\sin\theta)\cosh(\nu\theta)\,d\theta -\frac{2}{\pi}\sinh(\nu\pi/2)\int_0^{\infty}e^{-z\sinh t}\sin(\nu t)\,dt, \qquad |\arg z|<\pi/2, \tag{3.13}$$

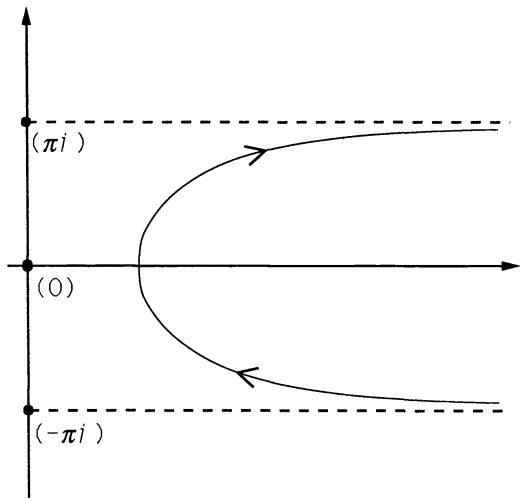


FIG. 1. *t plane.*

$$G_{i\nu}(z)=\frac{1}{\pi}\operatorname{cosech}(\nu\pi/2)\int_0^{\pi}\sin(z\sin\theta)\sinh(\nu\theta)\,d\theta$$

(3.14)

$$-\frac{2}{\pi}\cosh(\nu\pi/2)\int_0^{\infty}e^{-z\sinh t}\cos(\nu t)\,dt,\qquad |\arg z|<\pi/2.$$

Behavior at the singularities $z=0,\infty$. If ν (greater than zero) is fixed and $x\to 0^+$, then

$$F_{i\nu}(x)=\left(\frac{2\tanh(\nu\pi/2)}{\nu\pi}\right)^{1/2}\{\cos(\nu\ln(x/2)-\phi_{\nu,0})+O(x^2)\},\tag{3.15}$$

$$G_{i\nu}(x)=\left(\frac{2\coth(\nu\pi/2)}{\nu\pi}\right)^{1/2}\{\sin(\nu\ln(x/2)-\phi_{\nu,0})+O(x^2)\}.\tag{3.16}$$

Note that in a neighborhood of $x=0$ the amplitude of oscillation of $F_{i\nu}(x)$ and $G_{i\nu}(x)$ tends to 1 and ∞, respectively, as $\nu\to 0$.

As $z\to\infty$ in $|\arg z|\leqq\pi-\delta$

$$F_{i\nu}(z)\sim\left(\frac{2}{\pi z}\right)^{1/2}\left\{\cos(z-\pi/4)\sum_{s=0}^{\infty}(-1)^s\frac{A_{2s}(i\nu)}{z^{2s}}\right.$$

(3.17)

$$\left.-\sin(z-\pi/4)\sum_{s=0}^{\infty}(-1)^s\frac{A_{2s+1}(i\nu)}{z^{2s+1}}\right\},$$

$$G_{i\nu}(z)\sim\left(\frac{2}{\pi z}\right)^{1/2}\left\{\sin(z-\pi/4)\sum_{s=0}^{\infty}(-1)^s\frac{A_{2s}(i\nu)}{z^{2s}}\right.$$

(3.18)

$$\left.+\cos(z-\pi/4)\sum_{s=0}^{\infty}(-1)^s\frac{A_{2s+1}(i\nu)}{z^{2s+1}}\right\},$$

where the A_s are given by (4.02) of Chap. 7.

Zeros. From the foregoing results it is evident that $F_{i\nu}(x)$ and $G_{i\nu}(x)$ have an infinite number of zeros in both the x intervals $(0,\delta)$ and $[\delta,\infty)(\delta>0)$. A convenient notation for the zeros of $F_{i\nu}(x)$ is $\{f_{\nu,s}^{(<)}\}_{s=1}^{\infty}$ and $\{f_{\nu,s}^{(>)}\}_{s=1}^{\infty}$, where

$$\nu\tau>f_{\nu,1}^{(<)}>f_{\nu,2}^{(<)}>f_{\nu,3}^{(<)}>\cdots>0,\tag{3.19}$$

$$\nu\tau\leqq f_{\nu,1}^{(>)}<f_{\nu,2}^{(>)}<f_{\nu,3}^{(>)}<\cdots<\infty,\tag{3.20}$$

with τ being the positive constant defined by (5.6) below. Using the same convention we can denote the zeros of $G_{i\nu}(x)$ by $\{g_{\nu,s}^{(<)}\}_{s=1}^{\infty}$ and $\{g_{\nu,s}^{(>)}\}_{s=1}^{\infty}$. The zeros of $F_{i\nu}(x)$ and $G_{i\nu}(x)$ are simple and interlaced (cf. Lemma 1).

Asymptotic approximations for the zeros can be established in a similar manner to those derived in § 2 for the zeros of $K_{i\nu}(x)$ and $L_{i\nu}(x)$. For large ν the theory of § 8 of Chap. 6 (in particular § 8.5) can be applied to the uniform asymptotic expansions (5.15) and (5.16) (with $n=0$) which are given in § 5 below. We obtain the following asymptotic forms:

$$f_{\nu,s}^{(>)}=\nu Z\left\{\frac{(4s-1)\pi}{4\nu}\right\}+O\left(\frac{1}{\nu}\right)Z'\left\{\frac{(4s-1)}{4\nu}+O\left(\frac{1}{\nu^2}\right)\right\},\tag{3.21}$$

$$g_{\nu,s}^{(>)}=\nu Z\left\{\frac{(4s-3)\pi}{4\nu}\right\}+O\left(\frac{1}{\nu}\right)Z'\left\{\frac{(4s-3)}{4\nu}+O\left(\frac{1}{\nu^2}\right)\right\},\tag{3.22}$$

as $\nu\to\infty$, uniformly for all s. Here it is assumed that ν is sufficiently large to ensure that

$$|\varepsilon_{1,1}(\nu,0)|+|\varepsilon_{1,2}(\nu,0)|<\sqrt{2}.\tag{3.23}$$

This ensures that none of the zeros can take the value $\nu\tau$. In the Appendix we give a sufficient condition for (3.23) to hold, and it is seen that ν does not have to be very large for the inequality to hold (see (A6)). The function $Z(\xi)$ is defined implicitly by the equation

$$\xi = (1+Z^2)^{1/2} + \ln\left\{\frac{Z}{1+(1+Z^2)^{1/2}}\right\}. \tag{3.24}$$

The corresponding approximations for $f_{\nu,s}^{(<)}$ and $g_{\nu,s}^{(<)}$ are given by (3.21) and (3.22), respectively, with s replaced by $-s+1$.

For fixed s, (3.21) and (3.22) can be simplified by Taylor's theorem to give

$$f_{\nu,s}^{(>)} = \nu\tau + \frac{\tau(4s-1)\pi}{4(1+\tau^2)^{1/2}} + O\left(\frac{1}{\nu}\right), \tag{3.25}$$

$$g_{\nu,s}^{(>)} = \nu\tau + \frac{\tau(4s-3)\pi}{4(1+\tau^2)^{1/2}} + O\left(\frac{1}{\nu}\right), \tag{3.26}$$

as $\nu \to \infty$. Again, on replacing s by $-s+1$ in (3.25) and (3.26) we obtain the corresponding formulas for $f_{\nu,s}^{(<)}$ and $g_{\nu,s}^{(<)}$.

When ν is fixed, but still satisfying (3.23), and $x \to 0^+$ we employ the first two terms of (3.8) and (3.9) to obtain the approximations

$$f_{\nu,s}^{(>)} = 2\, e^{-(1/\nu)((s-1/2)\pi - \phi_{\nu,0})}\left\{1 - \frac{e^{-(2/\nu)((s-1/2)\pi-\phi_{\nu,0})}}{(1+\nu^2)} + O(e^{-(4\pi/\nu)(s-1/2)})\right\}, \tag{3.27}$$

$$g_{\nu,s}^{(>)} = 2\, e^{-(1/\nu)(s\pi - \phi_{\nu,0})}\left\{1 - \frac{e^{-(2/\nu)(s\pi-\phi_{\nu,0})}}{(1+\nu^2)} + O(e^{-(4s\pi/\nu)})\right\}, \tag{3.28}$$

as $s \to \infty$. Justification that these approximations represent the sth zero to the left of the point $x = \nu\tau$ follows in a similar manner to that of (2.26) and (2.32).

Finally, for fixed ν (satisfying (3.23)) and $s \to \infty$ we find from (3.17) that

$$f_{\nu,s}^{(>)} = \left(s - \frac{1}{4}\right)\pi + \frac{4\nu^2+1}{2(4s-1)\pi} - \frac{(4\nu^2+1)(28\nu^2+31)}{6(4s-1)^3\pi^3} + O(s^{-5}); \tag{3.29}$$

compare (6.03) of Chap. 7. Likewise, from (3.18) we find that

$$g_{\nu,s}^{(>)} = \left(s - \frac{3}{4}\right)\pi + \frac{4\nu^2+1}{2(4s-3)\pi} - \frac{(4\nu^2+1)(28\nu^2+31)}{6(4s-3)^3\pi^3} + O(s^{-5}). \tag{3.30}$$

4. Asymptotic expansions for modified Bessel functions of purely imaginary order. The modified Bessel functions $z^{1/2}K_{i\nu}(\nu z)$, $z^{1/2}L_{i\nu}(\nu z)$, as well as the analytic continuations $z^{1/2}K_{i\nu}(\nu z\, e^{\pi i})$, $z^{1/2}K_{i\nu}(\nu z\, e^{-\pi i})$, satisfy

$$\frac{d^2w}{dz^2} = \left\{\nu^2\frac{z^2-1}{z^2} - \frac{1}{4z^2}\right\}w, \tag{4.1}$$

which is characterized by having a regular singularity at $z=0$, an irregular singularity at $z=\infty$, and turning points at $z=\pm 1$. We apply the theory of a turning point in the complex plane (Theorem 9.1 of Chap. 11) to obtain asymptotic approximations, for large ν, in terms of Airy functions.

The first step is to transform (4.1) to the form

$$d^2W/d\zeta^2 = \{-\nu^2\zeta + \phi(\zeta)\}W, \tag{4.2}$$

which is achieved by the following Liouville transformation:

$$\frac{2}{3}\zeta^{3/2}(z)=\ln\left\{\frac{1+(1-z^2)^{1/2}}{z}\right\}-(1-z^2)^{1/2}, \tag{4.3}$$

$$W(\zeta)=\left(\frac{1-z^2}{z^2\zeta}\right)^{1/4}w(z). \tag{4.4}$$

This is precisely the Liouville transformation of § 10 of Chap. 11 and the reader is referred to this section for full details. It is seen from (10.04) in Chap. 11 that

$$\phi(\zeta)=\psi(-\zeta)=\frac{5}{16\zeta^2}-\frac{\zeta z^2(z^2+4)}{4(z^2-1)^3}. \tag{4.5}$$

In the notation of § 10 of Chap. 11 solutions of (4.2) are $W_{2n+1,0}(\nu,-\zeta)$, $W_{2n+1,1}(\nu,-\zeta)$, and $W_{2n+1,1}(\nu,-\zeta)$; see (9.02), (10.06), (10.07), (10.14), and (10.23) of Chap. 11. It can be shown by induction from (10.06) and (10.07) that

$$A_s(-\zeta)=(-1)^sA_s(\zeta),\qquad B_s(-\zeta)=(-1)^sB_s(\zeta), \tag{4.6}$$

and therefore for $j=0,\pm1$, $n=0,1,2,\cdots$, and $\nu>0$

$$\begin{aligned}W_{2n+1,j}(\nu,-\zeta)=&\operatorname{Ai}(-\nu^{2/3}\zeta e^{-2\pi ij/3})\sum_{s=0}^{n}(-1)^s\frac{A_s(\zeta)}{\nu^{2s}}\\&+\frac{\operatorname{Ai}'(-\nu^{2/3}\zeta e^{-2\pi ij/3})}{\nu^{4/3}}\sum_{s=0}^{n}(-1)^s\frac{B_s(\zeta)}{\nu^{2s}}+\varepsilon_{2n+1,j}(\nu,-\zeta).\end{aligned} \tag{4.7}$$

Bounds on the error terms $\varepsilon_{2n+1,j}$ are furnished by (9.03) of Chap. 11. The solutions above are to be identified with solutions of (4.1). First, since $z^{1/2}K_{i\nu}(\nu z)$ and $z^{1/2}(\zeta/(1-z^2))^{1/4}W_{2n+1,0}(\nu,-\zeta)$ are solutions of (4.1) that share the same recessive property at $z=+\infty$ ($\zeta=-\infty$), it follows that they are proportional to one another. The constant of proportionality can be determined by comparing the behavior of both functions at $z=\infty$, $\zeta=-\infty$; from (2.16), and from Chap. 11, (1.07), (10.08), (10.14), (10.23), we find

$$K_{i\nu}(\nu z)=\frac{\pi e^{-\nu\pi/2}}{\nu^{1/3}}\left(\frac{4\zeta}{1-z^2}\right)^{1/4}W_{2n+1,0}(\nu,-\zeta), \tag{4.8}$$

a result first given by Balogh (1967). (See also Exercise 10.6 of Chap. 11.)

The identification of $W_{2n+1,1}(\nu,-\zeta)$ is similar. Both this function, regarded as a function of z, and the modified Bessel function $K_{i\nu}(\nu z\,e^{-\pi i})$ are recessive at $z=\infty$ when $\pi/2<\arg z<\pi$. (We are restricting our attention to $|\arg z|<\pi$; $K_{i\nu}(\nu z\,e^{-\pi i})$ is of course also recessive at $z=\infty$ in $\pi\leqq\arg z<3\pi/2$.) It follows that there exists a constant c such that

$$K_{i\nu}(\nu z\,e^{-\pi i})=c\left(\frac{4\zeta}{1-z^2}\right)^{1/4}W_{2n+1,1}(\nu,-\zeta). \tag{4.9}$$

By comparing both sides as $\zeta\to\infty\ e^{-\pi i/3}$ we find that

$$c=\frac{\pi e^{\pi i/3}e^{\nu\pi/2}}{\nu^{1/3}} \tag{4.10}$$

Likewise, it can be shown that

$$K_{i\nu}(\nu z\,e^{\pi i})=\frac{\pi e^{-\pi i/3}e^{\nu\pi/2}}{\nu^{1/3}}\left(\frac{4\zeta}{1-z^2}\right)^{1/4}W_{2n+1,-1}(\nu,-\zeta). \tag{4.11}$$

This completes the identification of the asymptotic solutions (4.7). It remains to derive an asymptotic expansion for $L_{i\nu}(\nu z)$, and to do so we employ the analytic continuation formula (2.5a). On setting $m=\pm 1$ in this equation, and then eliminating $K_\mu(z)$ from the resulting two equations, we derive the relation

$$L_\mu(z)=\frac{1}{2\sin(\mu\pi)}\{K_\mu(z\,e^{-\pi i})-K_\mu(z\,e^{\pi i})\}. \tag{4.12}$$

We now replace z by νz in (4.12), set $\mu=i\nu$, and employ (4.9)-(4.11) to obtain the following identification:

$$\begin{aligned} L_{i\nu}(\nu z)=&\frac{\pi\, e^{\nu\pi/2}}{2i\nu^{1/3}\sinh(\nu\pi)}\left(\frac{4\zeta}{1-z^2}\right)^{1/4}\\ &\cdot\{e^{\pi i/3}W_{2n+1,1}(\nu,-\zeta)-e^{-\pi i/3}W_{2n+1,-1}(\nu,-\zeta)\}. \end{aligned} \tag{4.13}$$

On employing (4.7), together with (8.04) of Chap. 11, we can re-express this as

$$\begin{aligned} L_{i\nu}(\nu z)=&\frac{\pi\, e^{\nu\pi/2}}{2\nu^{1/3}\sinh(\nu\pi)}\left(\frac{4\zeta}{1-z^2}\right)^{1/4}\\ &\cdot\left[\mathrm{Bi}(-\nu^{2/3}\zeta)\sum_{s=0}^{n}(-1)^s\frac{A_s(\zeta)}{\nu^{2s}}+\frac{\mathrm{Bi}'(-\nu^{2/3}\zeta)}{\nu^{4/3}}\sum_{s=0}^{n-1}(-1)^s\frac{B_s(\zeta)}{\nu^{2s}}\right.\\ &\left.+\{e^{-\pi i/6}\varepsilon_{2n+1,1}(\nu,-\zeta)+e^{\pi i/6}\varepsilon_{2n+1,-1}(\nu,-\zeta)\}\right], \end{aligned} \tag{4.14}$$

an expansion that is uniformly valid for $\nu>0$ and $|\arg z|\leqq\pi-\delta$. We emphasize that both the expansions (4.8) and (4.14) are uniformly valid in a neighborhood of the singularity $z=0$, provided $|\arg z|\leqq\pi-\delta$. Use of (4.8) and (4.14) can be restricted to the half-plane $|\arg z|\leqq\pi/2$, extensions to other ranges of $\arg z$ being achieved via the analytic continuation formulas (2.5a, b).

An asymptotic expansion for $I_{i\nu}(\nu z)$ is also readily derived from the foregoing results; from the identity

$$I_{i\nu}(\nu z)=\frac{\operatorname{sech}(\nu\pi)}{2\pi i}\{e^{\nu\pi}K_{i\nu}(\nu z\,e^{-\pi i})-e^{-\nu\pi}K_{i\nu}(\nu z\,e^{\pi i})\},$$

and (4.9)-(4.11) we obtain

$$\begin{aligned} I_{i\nu}(\nu z)=&\frac{e^{\nu\pi/2}}{2\nu^{1/3}}\left(\frac{4\zeta}{1-z^2}\right)^{1/4}\left[\mathrm{Bi}_\nu(-\nu^{2/3})\sum_{s=0}^{n}(-1)^s\frac{A_s(\zeta)}{\nu^{2s}}\right.\\ &+\frac{\mathrm{Bi}'_\nu(-\nu^{2/3})}{\nu^{4/3}}\sum_{s=0}^{n-1}(-1)^s\frac{B_s(\zeta)}{\nu^{2s}}\\ &+\operatorname{sech}(\nu\pi)\{e^{\pi(\nu-i/6)}\varepsilon_{2n+1,1}(\nu,-\zeta)\\ &\left.+e^{-\pi(\nu-i/6)}\varepsilon_{2n+1,-1}(\nu,-\zeta)\}\right], \end{aligned} \tag{4.15}$$

where

$$\mathrm{Bi}_\nu(z)=\mathrm{Bi}(z)-i\tanh(\nu\pi)\,\mathrm{Ai}(z). \tag{4.16}$$

Again, this expansion is uniformly valid for $\nu>0$, $|\arg z|\leqq\pi-\delta$.

Finally, in this section we derive Debye-type expansions for $L_{i\nu}(\nu x)$, i.e., asymptotic expansions involving elementary functions that are uniformly valid for positive x. The corresponding Debye-type expansions for $K_{i\nu}(\nu x)$ are well known (see,

e.g., Magnus, Oberhettinger, and Soni (1966, p. 141)). The following expansions are not valid near the turning point $x=1$, and therefore we must consider the x intervals $(0, 1-\delta]$, $[1+\delta, \infty)$ separately.

First, consider the case $1+\delta \leqq x < \infty$. On applying the Liouville transformation to (4.1) (with z replaced by x), we obtain the transformed equation

$$d^2\Theta/d\eta^2 = \{\nu^2 + \chi(\eta)\}\Theta, \tag{4.17}$$

where

$$\eta(x) = \int_1^x \frac{(t^2-1)^{1/2}}{t}\,dt = (x^2-1)^{1/2} - \sec^{-1}(x), \tag{4.18}$$

$$\Theta(\eta) = (d\eta/dx)^{1/2} w(x) = x^{-1/2}(x^2-1)^{1/4} w(x), \tag{4.19}$$

$$\chi(\eta) = -\frac{x^2(4+x^2)}{4(x^2-1)}; \tag{4.20}$$

see § 2.1 of Chap. 10. On applying Theorem 3.1 of Chap. 10 we obtain the following solution of (4.17):

$$\Theta_n(\nu, \eta) = e^{-\nu\eta} \sum_{s=0}^{n-1} (-1)^s \frac{V_s(q)}{\nu^s} + \varepsilon_n(\nu, \eta), \tag{4.21}$$

where

$$q = (x^2-1)^{-1/2}, \tag{4.22}$$

$$V_0(q) = 1, \tag{4.23}$$

$$V_{s+1}(q) = \tfrac{1}{2}q^2(q^2+1)V_s'(q) + \tfrac{1}{8}\int_0^q V_s(t)(5t^2+1)\,dt \qquad (s \geqq 1). \tag{4.24}$$

A bound for $\varepsilon_n(\nu, \eta)$ is furnished by (3.04) of Chap. 10; for our purposes it suffices to observe that

$$\varepsilon_n(\nu, \eta) = e^{-\nu\eta} O(\nu^{-n})$$

as $\nu \to \infty$, uniformly for $1+\delta \leqq x < \infty$. It is possible to carry error bounds throughout the following analysis, but we will not pursue this.

Since both $K_{i\nu}(\nu x)$ and $\Theta_n(\nu, \eta)$ are recessive as $x \to \infty$ it follows that they are multiples of one another. By comparing both functions as $x \to \infty$ we find that

$$K_{i\nu}(\nu x) = (\pi/(2\nu))^{1/2}\, e^{-\nu\pi/2}(x^2-1)^{-1/4}\Theta_n(\nu, \eta). \tag{4.25}$$

Next, on identifying the left-hand side (LHS) of (4.25) with (4.8), employing asymptotic expansions for $\mathrm{Ai}(x)$, $\mathrm{Ai}'(x)$ of large positive argument (see Chap. 11, (1.07)), and equating coefficients of ν^{-s}, we arrive at the following relations for each $s \geqq 0$:

$$\sum_{j=0}^{s} (-1)^j \eta^{-2(s-j)} u_{2(s-j)} A_j(\zeta) + (-\zeta)^{1/2} \cdot \sum_{j=0}^{s-1} (-1)^j \eta^{-2(s-j)+1} v_{2(s-j)-1} B_j(\zeta) = V_{2s}(q), \tag{4.26}$$

$$\sum_{j=0}^{s} (-1)^j \eta^{-2(s-j)-1} u_{2(s-j)+1} A_j(\zeta) + (-\zeta)^{1/2} \cdot \sum_{j=0}^{s} (-1)^j \eta^{-2(s-j)} v_{2(s-j)} B_j(\zeta) = V_{2s+1}(q), \tag{4.27}$$

where $u_0 = v_0 = 1$ and

$$u_s = \frac{(2s+1)(2s+3)(2s+5)\cdots(6s-1)}{(216)^s s!}, \qquad v_s = -\frac{(6s+1)}{(6s-1)}u_s \qquad (s \geqq 1).$$

We now are in a position to derive a Debye-type asymptotic expansion for $L_{i\nu}(\nu x)$ for $x > 1$. From (4.14), (4.26), and (4.27), together with (1.07) and (1.16) of Chap. 11, we have for $1+\delta \leqq x < \infty$ and $\nu \to +\infty$:

$$L_{i\nu}(\nu x) \sim \left(\frac{\pi}{2\nu}\right)^{1/2} \frac{e^{\nu\pi/2}}{\sinh(\nu\pi)} (x^2-1)^{-1/4} e^{\nu\eta} \sum_{s=0}^{\infty} \frac{V_s(q)}{\nu^s}, \tag{4.28}$$

where $\eta(x)$, q, and $V_s(q)$ are given by (4.18) and (4.22)-(4.24). In a similar manner we can show that for $0 < x \leqq 1-\delta$, $\nu \to +\infty$,

$$\begin{aligned} L_{i\nu}(\nu x) \sim \left(\frac{\pi}{2\nu}\right)^{1/2} & \frac{e^{\nu\pi/2}}{\sinh(\nu\pi)} (1-x^2)^{-1/4} \\ & \cdot \left[-\sin(\nu\hat{\eta} - \pi/4) \sum_{s=0}^{n} \frac{V_{2s}(i\hat{q})}{\nu^{2s}} \right. \\ & \left. + \cos(\nu\hat{\eta} - \pi/4) \sum_{s=0}^{\infty} \frac{iV_{2s+1}(i\hat{q})}{\nu^{2s+1}} \right], \end{aligned} \tag{4.29}$$

where

$$\hat{q}(x) = (1-x^2)^{-1/2}, \tag{4.30}$$

$$\hat{\eta}(x) = \int_x^1 \frac{(1-t^2)^{1/2}}{t}\, dt = \ln\left\{\frac{1+(1-x^2)^{1/2}}{x}\right\} - (1-x^2)^{1/2}. \tag{4.31}$$

5. Asymptotic expansions for unmodified Bessel functions of purely imaginary order. The unmodified Bessel functions $z^{1/2}F_{i\nu}(\nu z)$, $z^{1/2}G_{i\nu}(\nu z)$, $z^{1/2}H_{i\nu}^{(1)}(\nu z)$, and $z^{1/2}H_{i\nu}^{(2)}(\nu z)$ satisfy the equation

$$\frac{d^2w}{dz^2} = \left\{-\nu^2 \frac{1+z^2}{z^2} - \frac{1}{4z^2}\right\} w, \tag{5.1}$$

which is characterized by having a regular singularity at $z = 0$, an irregular singularity at $z = \infty$, and turning points at $z = \pm i$ (where the results of § 4 are applicable).

We restrict our attention to the half-plane $|\arg z| < \pi/2$, with $\nu > 0$, and apply the Liouville transformation of § 7 in Chap. 10. The effect of this transformation is to throw (5.1) into the form

$$d^2W/d\xi^2 = \{-\nu^2 + \psi(\xi)\} W, \tag{5.2}$$

where

$$\xi(z) = (1+z^2)^{1/2} + \ln\left\{\frac{z}{1+(1+z^2)^{1/2}}\right\}, \tag{5.3}$$

$$W(\xi) = \left(\frac{1+z^2}{z^2}\right)^{1/4} w(z), \tag{5.4}$$

$$\psi(\xi) = \frac{z^2(4-z^2)}{4(1+z^2)^3}. \tag{5.5}$$

Before proceeding further let us introduce a constant τ, defined by

$$\tau = (\tau_0 - 1)^{1/2}, \tag{5.6}$$

where τ_0 is the positive root of the equation $\coth \tau_0 = \tau_0$; from Exercise 8.1 of Chap. 10 and (5.3) it is seen that $z = \tau$ is the point that is mapped to $\xi = 0$, i.e.,

$$\xi(\tau) = 0 \qquad (\tau = 0.6627\cdots). \tag{5.7}$$

On applying Theorem 3.1 of Chap. 10 to the transformed equation (5.2) we obtain the following solutions:

$$W_{n,1}(\nu, \xi) = e^{i\nu\xi} \sum_{s=0}^{n-1} \frac{U_s(p)}{(i\nu)^s} + \varepsilon_{n,1}(\nu, \xi), \tag{5.8}$$

$$W_{n,2}(\nu, \xi) = e^{-i\nu\xi} \sum_{s=0}^{n-1} \frac{(-1)^s U_s(p)}{(i\nu)^s} + \varepsilon_{n,2}(\nu, \xi), \tag{5.9}$$

where

$$p = (1+z^2)^{-1/2}, \tag{5.10}$$

and the coefficients $U_s(p)$ are given by (7.10) of Chap. 10, and are related to the V_s of the previous section by

$$U_s(p) = (-i)^s V_s(ip) \qquad (s = 0, 1, 2, \cdots). \tag{5.11}$$

Our choice of reference points for the solutions is $\alpha_1 = +i\infty$, $\alpha_2 = -i\infty$; with these choices the error term $\varepsilon_{n,1}$ is bounded by (3.04) of Chap. 10 for all points in $|\arg z| \leqq \pi/2$ except those on the finite interval $z = i\sigma$, $0 \leqq \sigma \leqq 1$, and at $z = -i$, with the corresponding bound for $\varepsilon_{n,2}$ being valid in the conjugate region.

We now identify the solutions (5.8) and (5.9) with Bessel functions. First, we see that for some constant c_1

$$H_{i\nu}^{(1)}(\nu z) = c_1 (1+z^2)^{-1/4} W_{n,1}(\nu, \xi), \tag{5.12}$$

since both functions are solutions of Bessel's equation and share the same recessive property as $z \to +i\infty$. By comparing both sides as $z \to +i\infty$ (see (4.03) of Chap. 7) we find that

$$c_1 = (2/(\pi\nu))^{1/2}\, e^{\nu\pi/2}\, e^{-i\pi/4}. \tag{5.13}$$

Likewise we find that

$$H_{i\nu}^{(2)}(\nu z) = (2/(\pi\nu))^{1/2}\, e^{-\nu\pi/2}\, e^{i\pi/4} (1+z^2)^{-1/4} W_{n,2}(\nu, \xi). \tag{5.14}$$

Asymptotic expansions for $F_{i\nu}(\nu z)$ and $G_{i\nu}(\nu z)$ are now obtainable from the above expansions and the relations (3.2) and (3.3): for $\nu > 0$, $|\arg z| < \pi/2$ we have

$$\begin{aligned} F_{i\nu}(\nu z) = {} & \left(\frac{2}{\pi\nu}\right)^{1/2} (1+z^2)^{-1/4} \\ & \cdot \Bigg[\cos(\nu\xi - \pi/4) \sum_{s=0}^{n} \frac{(-1)^s U_{2s}(p)}{\nu^{2s}} \\ & + \sin(\nu\xi - \pi/4) \sum_{s=0}^{n-1} \frac{(-1)^s U_{2s+1}(p)}{\nu^{2s+1}} \\ & + \frac{1}{2}\{e^{-i\pi/4} \varepsilon_{2n+1,1}(\nu, \xi) + e^{i\pi/4} \varepsilon_{2n+1,2}(\nu, \xi)\} \Bigg], \end{aligned} \tag{5.15}$$

$$G_{i\nu}(\nu z)=\left(\frac{2}{\pi\nu}\right)^{1/2}(1+z^2)^{-1/4}$$
$$\cdot\Bigg[\sin(\nu\xi-\pi/4)\sum_{s=1}^{n}\frac{(-1)^s U_{2s}(p)}{\nu^{2s}}$$
$$-\cos(\nu\xi-\pi/4)\sum_{s=0}^{n-1}\frac{(-1)^s U_{2s+1}(p)}{\nu^{2s+1}} \tag{5.16}$$
$$+\frac{1}{2i}\{e^{-i\pi/4}\varepsilon_{2n+1,1}(\nu,\xi)-e^{i\pi/4}\varepsilon_{2n+1,2}(\nu,\xi)\}\Bigg].$$

Asymptotic expansions for $J_{i\nu}(\nu z)$ and $Y_{i\nu}(\nu z)$ can also be obtained in a similar manner. Note that the above expansions are uniformly valid in a neighborhood of $z=0$, provided $-\pi/2\leqq\arg z\leqq\pi/2-\delta$ for (5.12), $-\pi/2+\delta\leqq\arg z\leqq\pi/2$ for (5.14), and $|\arg z|\leqq\pi/2-\delta$ for both (5.15) and (5.16).

6. Auxiliary functions. For differential equations of the type (1.1), with (1.2) applying, asymptotic solutions will be obtained involving Bessel functions and modified Bessel functions of purely imaginary order. In order to construct error bounds it is necessary to define auxiliary weight, modulus, and phase functions for these functions, as Olver did for the corresponding problem of Chap. 12 (see § 1.3).

First we define auxiliary functions for $K_{i\nu}(x)$ and $L_{i\nu}(x)$. *Let* $x=\chi_\nu$ be the largest positive root of

$$K_{i\nu}(x)-L_{i\nu}(x)=0. \tag{6.1}$$

Since the LHS of (6.1) is negative as $x\to\infty$, and positive at $x=l_{\nu,1}$ (see (2.16), (2.17), and (2.20)), it follows that

$$l_{\nu,1}<\chi_\nu<\infty. \tag{6.2}$$

We now define a weight function $E_\nu^{(1)}(x)$ by

$$E_\nu^{(1)}(x)=1 \qquad (0\leqq x\leqq\chi_\nu), \tag{6.3a}$$

$$E_\nu^{(1)}(x)=\left\{\frac{L_{i\nu}(x)}{K_{i\nu}(x)}\right\}^{1/2} \qquad (\chi_\nu\leqq x<\infty). \tag{6.3b}$$

From the definition above it is seen that $E_\nu^{(1)}(x)$ is a positive continuous function of x, and moreover is nondecreasing, as can be seen from the equation

$$\frac{d\{E_\nu^{(1)}(x)\}^2}{dx}=\frac{\pi}{\sinh(\nu\pi)xK_{i\nu}^2(x)} \qquad (\chi_\nu<x<\infty), \tag{6.4}$$

which can be derived by differentiating (6.3b) and employing (2.10).

Having defined a weight function we now introduce modulus and phase functions; we define them by the relations

$$K_{i\nu}(x)=-(E_\nu^{(1)}(x))^{-1}M_\nu^{(1)}(x)\sin\theta_\nu^{(1)}(x), \tag{6.5a}$$

$$L_{i\nu}(x)=E_\nu^{(1)}(x)M_\nu^{(1)}(x)\cos\theta_\nu^{(1)}(x), \tag{6.5b}$$

or explicitly

$$M_\nu^{(1)}(x)=\{K_{i\nu}^2(x)+L_{i\nu}^2(x)\}^{1/2},$$
$$\theta_\nu^{(1)}(x)=-\tan^{-1}\{K_{i\nu}(x)/L_{i\nu}(x)\} \qquad (0<x\leqq\chi_\nu), \tag{6.6a}$$

(6.6b)
$$M_\nu^{(1)}(x)=\{2K_{i\nu}(x)L_{i\nu}(x)\}^{1/2},$$
$$\theta_\nu^{(1)}(x)=-\frac{\pi}{4} \qquad (\chi_\nu \leqq x<\infty),$$

the branch of the inverse tangent being chosen so that $\theta_\nu(x)$ is continuous for $0<x<\infty$. On differentiating and using (2.10) we find that

(6.7)
$$\frac{d\theta_\nu^{(1)}(x)}{dx}=\frac{\pi}{\sinh(\nu\pi)x(M_\nu^{(1)}(x))^2}>0 \qquad (0<x\leqq\chi_\nu)$$

and therefore as x decreases from χ_ν to zero, $\theta_\nu^{(1)}(x)$ decreases monotonically from $-\pi/4$ to $-\infty$. This fact, together with (2.20), (6.2), and (6.5a, b), shows that

(6.8)
$$\theta_\nu^{(1)}(k_{\nu,s})=-s\pi, \qquad \theta_\nu^{(1)}(l_{\nu,s})=-(s-\tfrac{1}{2})\pi.$$

For fixed $\nu>0$, the following asymptotic behavior of the auxiliary functions can readily be derived from the definitions above and the results of § 2.

As $x\to 0^+$

(6.9)
$$M_\nu^{(1)}(x)\to\left(\frac{\pi}{\nu\sinh(\nu\pi)}\right)^{1/2}, \qquad \theta_\nu^{(1)}(x)\sim\nu\ln\left(\frac{x}{2}\right);$$

as $x\to\infty$

(6.10)
$$E_\nu^{(1)}(x)\sim\frac{e^x}{(\sinh(\nu\pi))^{1/2}}, \qquad M_\nu^{(1)}(x)\sim\left(\frac{\pi}{x\sinh(\nu\pi)}\right)^{1/2}.$$

Next, we must introduce auxiliary functions for the derivatives of the modified Bessel functions. We define

(6.11)
$$K'_{i\nu}(x)=(E_\nu^{(1)}(x))^{-1}N_\nu^{(1)}(x)\sin\omega_\nu^{(1)}(x),$$

(6.12)
$$L'_{i\nu}(x)=E_\nu^{(1)}(x)N_\nu^{(1)}(x)\cos\omega_\nu^{(1)}(x),$$

giving

(6.13a)
$$N_\nu^{(1)}(x)=\{K'^2_{i\nu}(x)+L'^2_{i\nu}(x)\}^{1/2}$$

(6.13b)
$$\omega_\nu^{(1)}(x)=\tan^{-1}\{K'_{i\nu}(x)/L'_{i\nu}(x)\} \qquad (0<x\leqq\chi_\nu),$$

(6.14a)
$$N_\nu^{(1)}(x)=\left(\frac{K'^2_{i\nu}(x)L^2_{i\nu}(x)+L'^2_{i\nu}(x)K^2_{i\nu}(x)}{K_{i\nu}(x)L_{i\nu}(x)}\right)^{1/2}$$

(6.14b)
$$\omega_\nu^{(1)}(x)=\tan^{-1}\left(\frac{K'_{i\nu}(x)L_{i\nu}(x)}{L'_{i\nu}(x)K_{i\nu}(x)}\right) \qquad (\chi_\nu\leqq x<\infty).$$

The branches of the inverse tangents are chosen so that $\omega_\nu^{(1)}(x)$ is continuous for $0<x<\infty$ with $\omega_\nu^{(1)}(x)\to-\pi/4$ as $x\to\infty$. From the following equation (which can be deduced from (2.1), (2.10), and (6.11)-(6.13)):

$$\frac{d\omega_\nu^{(1)}(x)}{dx}=\frac{\pi(x^2-\nu^2)}{\sinh(\nu\pi)x^3(N_\nu^{(1)}(x))^2} \qquad (0<x\leqq\chi_\nu),$$

it is seen that $\omega_\nu^{(1)}(x)$ is monotonically decreasing for $0<x<\hat{\chi}_\nu$, where $\hat{\chi}_\nu=\min(\nu,\chi_\nu)$.

The asymptotic behavior of $N_\nu^{(1)}(x)$ is as follows. As $x\to 0^+$

(6.15)
$$N_\nu^{(1)}(x)\sim\frac{2}{x}\left(\frac{\nu\pi}{\sinh(\nu\pi)}\right)^{1/2}.$$

As $x\to\infty$

(6.16)
$$N_\nu^{(1)}(x)\sim\left(\frac{\pi}{x\sinh(\nu\pi)}\right)^{1/2}.$$

Auxiliary functions for the unmodified Bessel functions $F_{i\nu}(x)$ and $G_{i\nu}(x)$ are defined in a similar manner. These functions are oscillatory, of bounded amplitude, throughout the x interval $(0,\infty)$, and as such a weight function does not strictly need to be introduced. However, although the amplitudes of both functions are equal for large ν, this is not the case when ν is small; near the origin the amplitudes of the two functions are quite disparate as $\nu\to 0$ (see (3.15) and (3.16)). Thus, to sharpen subsequent error bounds, we introduce a weight function $E_\nu^{(2)}(x)$ for $G_{i\nu}(x)$ that is continuous in x and decreases monotonically from the value $\coth(\nu\pi/2)$ at $x=0$ to unity at $x=\infty$. Our choice, one of the simplest, is

$$E_\nu^{(2)}(x)=\frac{1+x}{\tanh(\nu\pi/2)+x}. \tag{6.17}$$

We now introduce modulus and phase functions in the usual manner. We define

$$F_{i\nu}(x)=M_\nu^{(2)}(x)\cos\theta_\nu^{(2)}(x), \tag{6.18}$$

$$G_{i\nu}(x)=E_\nu^{(2)}(x)M_\nu^{(2)}(x)\sin\theta_\nu^{(2)}(x), \tag{6.19}$$

so that

$$M_\nu^{(2)}(x)=\left\{F_{i\nu}^2(x)+\left(\frac{G_{i\nu}(x)}{E_\nu^{(2)}(x)}\right)^2\right\}^{1/2}, \tag{6.20}$$

$$\theta_\nu^{(2)}(x)=\tan^{-1}\left\{\frac{G_{i\nu}(x)}{E_\nu^{(2)}(x)F_{i\nu}(x)}\right\}. \tag{6.21}$$

On differentiating (6.21) and employing (3.10) we arrive at the equation

$$E_\nu^{(2)}(x)(M_\nu^{(2)}(x))^2\frac{d\theta_\nu^{(2)}(x)}{dx}=\frac{2}{\pi x}+F_{i\nu}(x)G_{i\nu}(x)\frac{1-\tanh(\nu\pi/2)}{(\tanh(\nu\pi/2)+x)(1+x)}.$$

From (3.15)-(3.18), therefore, it is seen that $d\theta_\nu^{(2)}(x)/dx$ is positive for both sufficiently small and sufficiently large x, for each fixed positive value of ν. $\theta_\nu^{(2)}(x)$ is thus monotonically increasing for large x, and with this fact in mind we define the branch of the inverse tangent in (6.21) so that $\theta_\nu^{(2)}(x)$ is continuous for $0<x<\infty$, and also

$$\theta_\nu^{(2)}(x)=x-\frac{\pi}{4}+o(1)\quad\text{as } x\to\infty. \tag{6.22}$$

The asymptotic behavior of $M_\nu^{(2)}(x)$ is as follows. As $x\to 0^+$

$$M_\nu^{(2)}(x)\to\left(\frac{2\tanh(\nu\pi/2)}{\nu\pi}\right)^{1/2}. \tag{6.23}$$

As $x\to\infty$

$$M_\nu^{(2)}(x)\sim\left(\frac{2}{\pi x}\right)^{1/2}. \tag{6.24}$$

Finally, we define modulus and phase functions for the derivatives of $F_{i\nu}(x)$ and $G_{i\nu}(x)$ by

$$F'_{i\nu}(x)=N_\nu^{(2)}(x)\cos\omega_\nu^{(2)}(x), \tag{6.25}$$

$$G'_{i\nu}(x)=E_\nu^{(2)}(x)N_\nu^{(2)}(x)\sin\omega_\nu^{(2)}(x), \tag{6.26}$$

or explicitly

$$N_\nu^{(2)}(x)=\left\{F_{i\nu}'^2(x)+\left(\frac{G'_{i\nu}(x)}{E_\nu^{(2)}(x)}\right)^2\right\}^{1/2}, \tag{6.27}$$

$$\omega_\nu^{(2)}(x)=\tan^{-1}\left\{\frac{G'_{i\nu}(x)}{E_\nu^{(2)}(x)F'_{i\nu}(x)}\right\}. \tag{6.28}$$

The derivative of (6.28) can be shown to be

$$\begin{aligned} &x^2E_\nu^{(2)}(x)(N_\nu^{(2)}(x))^2\frac{d\omega_\nu^{(2)}(x)}{dx} \\ &=\frac{2(x^2+\nu^2)}{\pi x}+\frac{x^2F'_{i\nu}(x)G'_{i\nu}(x)(1-\tanh(\nu\pi))}{(\tanh(\nu\pi/2)+x)(1+x)}, \end{aligned} \tag{6.29}$$

and since the product $F'_{i\nu}(x)G'_{i\nu}(x)$ is $O(1/x)$ as $x\to\infty$ it follows that $d\omega_\nu^{(2)}(x)/dx$ is positive for sufficiently large x. The function $\omega_\nu^{(2)}(x)$ is thus monotonically increasing as $x\to\infty$, and therefore we can define the branch of the inverse tangent of (6.28) so that $\omega_\nu^{(2)}(x)$ is continuous for $0<x<\infty$, with the stipulation that

$$\omega_\nu^{(2)}(x)=x+\frac{\pi}{4}+o(1)\quad\text{as } x\to\infty. \tag{6.30}$$

The asymptotic forms of $N_\nu^{(2)}(x)$ are

$$N_\nu^{(2)}(x)\sim\frac{2}{x}\left(\frac{2\nu\tanh(\nu\pi/2)}{\pi}\right)^{1/2}\quad\text{as } x\to0^+, \tag{6.31}$$

$$N_\nu^{(2)}(x)\sim\left(\frac{2}{\pi x}\right)^{1/2}\quad\text{as } x\to\infty. \tag{6.32}$$

7. Asymptotic expansions for solutions of a differential equation with a large parameter and a simple pole. We now turn our attention to differential equations of the form

$$\frac{d^2W}{d\zeta^2}=\left\{\frac{u^2}{4\zeta}-\frac{\nu^2+1}{4\zeta^2}+\frac{\psi(\zeta)}{\zeta}\right\}W, \tag{7.1}$$

where u and ν are positive parameters, and $\psi(\zeta)$ is analytic in some interval $[0,\beta)$, where β is positive (and possibly infinite). Our task is to construct asymptotic solutions for (7.1) for large u, analogous to those of Olver, who in Chap. 12 considered the complementary problem where the coefficient of ζ^{-2} is greater than or equal to $-\frac{1}{4}$.

We start by considering the comparison equation to (7.1):

$$\frac{d^2W}{d\zeta^2}=\left\{\frac{u^2}{4\zeta}-\frac{\nu^2+1}{4\zeta^2}\right\}W, \tag{7.2}$$

which has exact solutions $\zeta^{1/2}\mathscr{L}_{i\nu}(u\zeta^{1/2})$, where $\mathscr{L}_{i\nu}$ denotes $K_{i\nu}$ or $L_{i\nu}$, or any linear combination of the two. These solutions are in fact the first terms in asymptotic expansions of solutions of (7.1); we seek formal series solutions of the form

$$W=\zeta^{1/2}\mathscr{L}_{i\nu}(u\zeta^{1/2})\sum_{s=0}^{\infty}\frac{C_s(\zeta)}{u^{2s}}+\frac{\zeta}{u}\mathscr{L}'_{i\nu}(u\zeta^{1/2})\sum_{s=0}^{\infty}\frac{D_s(\zeta)}{u^{2s}}\qquad(\zeta>0). \tag{7.3}$$

On substituting the series above into (7.1) and comparing like powers of u we find that the series formally satisfies the equation if both

$$\zeta C''_s(\zeta)+C'_s(\zeta)-\psi(\zeta)C_s(\zeta)+D'_s(\zeta)-\nu^2D'_{s-1}(\zeta)+\tfrac{1}{2}D_s(\zeta)=0, \tag{7.4}$$

$$C'_{s+1}(\zeta)+\zeta D''_s(\zeta)+D'_s(\zeta)-\psi(\zeta)D_s(\zeta)=0. \tag{7.5}$$

These two equations can be integrated to give the following two recursion relations for the coefficients:

$$D_s(\zeta)=-C'_s(\zeta)+\zeta^{-1/2}\int_0^\zeta t^{-1/2}\{\psi(t)C_s(t)-C'_s(t)/2+\nu^2D'_{s-1}(t)\}\,dt, \tag{7.6}$$

$$C_{s+1}(\zeta) = -\zeta D_s'(\zeta) + \int \psi(\zeta) D_s(\zeta)\, d\zeta. \tag{7.7}$$

Without loss of generality we set $C_0(\zeta) = 1$. Since, by hypotheses, $\psi(\zeta)$ is analytic, so too are the coefficients C_s, D_s in the ζ interval $[0, \beta)$ (see Chap. 11, Lemma 7.1).

Before we state our theorem on error bounds, let us define certain constants that appear:

$$\lambda_1^{(1)}(\nu) = \sup\{\sigma(\nu) x (M_\nu^{(1)}(x))^2\}, \tag{7.8}$$

$$\lambda_2^{(1)}(\nu) = \sup\{\sigma(\nu) x |K_{i\nu}(x)| E_\nu^{(1)}(x) M_\nu^{(1)}(x)\}, \tag{7.9}$$

$$\lambda_3^{(1)}(\nu) = \sup\{\sigma(\nu) x |L_{i\nu}(x)| (E_\nu^{(1)}(x))^{-1} M_\nu^{(1)}(x)\}, \tag{7.10}$$

where

$$\sigma(\nu) = 2 \sinh(\nu\pi)/\pi,$$

each supremum being taken over the finite x interval $(0, \infty)$. It is readily shown that each supremum exists and is finite for every positive value of ν.

THEOREM 1. *With the conditions stated at the beginning of this section,* (7.1) *has, for each positive value of* u *and* ν *and each integer* n, *the following two solutions, which are repeatedly differentiable in the* ζ *interval* $(0, \beta)$:

$$W_{2n+1,1}(u, \zeta) = \zeta^{1/2} K_{i\nu}(u\zeta^{1/2}) \sum_{s=0}^{n} \frac{C_s(\zeta)}{u^{2s}} + \frac{\zeta}{u} K_{i\nu}'(u\zeta^{1/2}) \cdot \sum_{s=0}^{n-1} \frac{D_s(\zeta)}{u^{2s}} + \varepsilon_{2n+1,1}(u, \zeta), \tag{7.11}$$

$$W_{2n+1,2}(u, \zeta) = \zeta^{1/2} L_{i\nu}(u\zeta^{1/2}) \sum_{s=0}^{n} \frac{C_s(\zeta)}{u^{2s}} + \frac{\zeta}{u} L_{i\nu}'(u\zeta^{1/2}) \cdot \sum_{s=0}^{n-1} \frac{D_s(\zeta)}{u^{2s}} + \varepsilon_{2n+1,2}(u, \zeta), \tag{7.12}$$

where

$$\frac{|\varepsilon_{2n+1,1}(u, \zeta)|}{\zeta^{1/2} M_\nu^{(1)}(u\zeta^{1/2})}, \frac{|\partial \varepsilon_{2n+1,1}(u, \zeta)/\partial\zeta|}{\{\zeta^{-1/2} M_\nu^{(1)}(u\zeta^{1/2}) + u N_\nu^{(1)}(u\zeta^{1/2})\}/2} \leqq \lambda_2^{(1)}(\nu)(E_\nu^{(1)}(u\zeta^{1/2}))^{-1} \exp\left\{\frac{\lambda_1^{(1)}(\nu)}{u} \mathscr{V}_{\zeta,\beta}(\zeta^{1/2} D_0(\zeta))\right\} \cdot \frac{\mathscr{V}_{\zeta,\beta}(\zeta^{1/2} D_n(\zeta))}{u^{2n+1}}, \tag{7.13}$$

$$\frac{|\varepsilon_{2n+1,2}(u, \zeta)|}{\zeta^{1/2} M_\nu^{(1)}(u\zeta^{1/2})}, \frac{|\partial \varepsilon_{2n+1,2}(u, \zeta)/\partial\zeta|}{\{\zeta^{-1/2} M_\nu^{(1)}(u\zeta^{1/2}) + u N_\nu^{(1)}(u\zeta^{1/2})\}/2} \leqq \lambda_3^{(1)}(\nu) E_\nu^{(1)}(u\zeta^{1/2}) \exp\left\{\frac{\lambda_1^{(1)}(\nu)}{u} \mathscr{V}_{0,\zeta}(\zeta^{1/2} D_0(\zeta))\right\} \cdot \frac{\mathscr{V}_{0,\zeta}(\zeta^{1/2} D_n(\zeta))}{u^{2n+1}}. \tag{7.14}$$

The derivation of these error bounds is similar to that of Theorem 4.1 of Chap. 12, and details will not be included here.

It remains to construct asymptotic solutions for (7.1) in the interval $(\alpha, 0)$, where α is negative (possibly infinite) constant. On replacing ζ by $|\zeta|e^{\pi i}$ in (7.3) it is readily verified that formal series solutions of (7.1) are given by

$$W = |\zeta|^{1/2}\mathscr{C}_{i\nu}(u|\zeta|^{1/2}) \sum_{s=0}^{\infty} \frac{C_s(\zeta)}{u^{2s}} + \frac{|\zeta|}{u}\mathscr{C}'_{i\nu}(u|\zeta|^{1/2}) \sum_{s=0}^{\infty} \frac{D_s(\zeta)}{u^{2s}} \qquad (\zeta<0), \tag{7.15}$$

where $\mathscr{C}_{i\nu}$ denotes $F_{i\nu}$ or $G_{i\nu}$, or any linear combination of the two. The coefficients $C_s(\zeta)$ and $D_s(\zeta)$ here are understood to be the analytic continuations across $\zeta = 0$ of those satisfying (7.4) and (7.5); thus (7.7) still holds, and (7.6) is replaced by

$$D_s(\zeta) = -C'_s(\zeta) + |\zeta|^{-1/2}\int_{\zeta}^{0} |t|^{-1/2}\{\psi(t)C_s(t) - C'_s(t)/2 + \nu^2 D'_{s-1}(t)\}\, dt. \tag{7.16}$$

As before, we introduce three constants that appear in subsequent error bounds. We define

$$\lambda_1^{(2)}(\nu) = \sup\{\pi x E_\nu^{(2)}(x)(M_\nu^{(2)}(x))^2\}, \tag{7.17}$$

$$\lambda_2^{(2)}(\nu) = \sup\{\pi x |F_{i\nu}(x)| E_\nu^{(2)}(x) M_\nu^{(2)}(x)\}, \tag{7.18}$$

$$\lambda_3^{(2)}(\nu) = \sup\{\pi x |G_{i\nu}(x)| M_\nu^{(2)}(x)\}, \tag{7.19}$$

each supremum being taken over the x interval $(0, \infty)$; again, it is readily verified that each supremum exists and is finite for every positive value of ν.

We may now state the following theorem concerning error bounds.

THEOREM 2. *With the conditions stated at the beginning of this section,* (7.1) *has, for each positive value of u and ν and each integer n, the following two solutions, which are repeatedly differentiable in the ζ interval* $(\alpha, 0)$:

$$W_{2n+1,3}(u,\zeta) = |\zeta|^{1/2}F_{i\nu}(u|\zeta|^{1/2}) \sum_{s=0}^{n} \frac{C_s(\zeta)}{u^{2s}} + \frac{|\zeta|}{u}F'_{i\nu}(u|\zeta|^{1/2}) \sum_{s=0}^{n-1} \frac{D_s(\zeta)}{u^{2s}} + \varepsilon_{2n+1,3}(u,\zeta), \tag{7.20}$$

$$W_{2n+1,4}(u,\zeta) = |\zeta|^{1/2}G_{i\nu}(u|\zeta|^{1/2}) \sum_{s=0}^{n} \frac{C_s(\zeta)}{u^{2s}} + \frac{|\zeta|}{u}G'_{i\nu}(u|\zeta|^{1/2}) \sum_{s=0}^{n-1} \frac{D_s(\zeta)}{u^{2s}} + \varepsilon_{2n+1,4}(u,\zeta), \tag{7.21}$$

where

$$\frac{|\varepsilon_{2n+1,3}(u,\zeta)|}{|\zeta|^{1/2}M_\nu^{(2)}(u|\zeta|^{1/2})}, \frac{|\partial\varepsilon_{2n+1,3}(u,\zeta)/\partial\zeta|}{\{|\zeta|^{-1/2}M_\nu^{(2)}(u|\zeta|^{1/2}) + uN_\nu^{(2)}(u|\zeta|^{1/2})\}/2}$$

$$\leqq \lambda_2^{(2)}(\nu)\exp\left\{\frac{\lambda_1^{(2)}(\nu)}{u}\mathscr{V}_{\zeta,0}(|\zeta|^{1/2}D_0(\zeta))\right\} \cdot \frac{\mathscr{V}_{\zeta,0}(|\zeta|^{1/2}D_n(\zeta))}{u^{2n+1}}, \tag{7.22}$$

$$\frac{|\varepsilon_{2n+1,4}(u,\zeta)|}{|\zeta|^{1/2}M_\nu^{(2)}(u|\zeta|^{1/2})},\frac{|\partial\varepsilon_{2n+1,4}(u,\zeta)/\partial\zeta|}{\{|\zeta|^{-1/2}M_\nu^{(2)}(u|\zeta|^{1/2})+uN_\nu^{(2)}(u|\zeta|^{1/2})\}/2}$$

$$\leqq\lambda_3^{(2)}(\nu)E_\nu^{(2)}(u|\zeta|^{1/2})\exp\left\{\frac{\lambda_1^{(2)}(\nu)}{u}\mathcal{V}_{\alpha,\zeta}(|\zeta|^{1/2}D_0(\zeta))\right\} \tag{7.23}$$

$$\cdot\frac{\mathcal{V}_{\alpha,\zeta}(|\zeta|^{1/2}D_n(\xi))}{u^{2n+1}}.$$

Final remarks. (i) We have assumed that $\psi(\zeta)$ is infinitely differentiable in (α,β). If we do not require asymptotic expansions for solutions of (7.1), but just a finite number of terms in the approximations, the requirement of analyticity of $\psi(\zeta)$ can be relaxed to that of finite differentiability.

(ii) Since we have derived explicit error bounds on the approximations, it has not been necessary to impose any restrictions on the dependence of $\psi(\zeta)$ on u, other than that it be a continuous function of u; if the dependence of ψ on u adversely affects the asymptotic validity of an approximation it will be reflected in the error bound.

(iii) To facilitate identification of solutions it is desirable that the asymptotic solutions be uniformly valid on the semi-infinite ζ intervals $(-\infty,0)$ and $(0,\infty)$. For this it is necessary that the variations of $|\zeta|^{1/2}D_s(\zeta)$ $(s=0,1,2,\cdots,n)$ converge at $\zeta=\pm\infty$. Sufficient conditions for this to be true are given in Exercise 4.2 of Chap. 12.

(iv) The error bounds can be used to deduce the asymptotic behavior of the four solutions, both respect to the independent variable ζ and the asymptotic variable u. For instance the solution $W_{2n+1,1}(u,\zeta)$ is seen to be recessive as $\zeta\to\beta$, a property that uniquely characterizes the solution if $\beta=\infty$. Likewise, the solutions $W_{2n+1,2}(u,\zeta)$ and $W_{2n+1,3}(u,\zeta)$ can be identified by their behavior as $\zeta\to0$ (with the aid of (2.15) and (3.15)), and $W_{2n+1,4}(u,\zeta)$ can be identified by its behavior as $\zeta\to\alpha$.

Finally, consider the asymptotic behavior of the four solutions as $u\to\infty$. If the variations in the error bounds are bounded functions of u then, in the manner of § 5.2 of Chap. 12, it can be shown that the RHS of (7.3) provides a uniform compound expansion of $W_{2n+1,1}(u,\zeta)$ and $W_{2n+1,2}(u,\zeta)$, for $\mathscr{L}=K$ and L, respectively, to $2n+1$ terms. A similar argument holds for (7.15). The existence of solutions that are independent of n and have the infinite series (7.3) or (7.15) as compound asymptotic expansions may be established by the method of § 6 of Chap. 10.

Appendix. We investigate how large ν should be to ensure that (3.23) holds. First, we observe that $\varepsilon_{1,1}(\nu,0)$ is bounded by

$$|\varepsilon_{1,1}(\nu,0)\leqq2\exp\left\{\frac{2\mathcal{V}_\Lambda(U_1)}{\nu}\right\}\frac{\mathcal{V}_\Lambda(U_1)}{\nu}, \tag{A.1}$$

where

$$U_1(p)=(3p-5p^3)/24, \tag{A.2}$$

and p is given by (5.10). The bound (A.1) corresponds to (7.14) of Chap. 10, the only difference being that the reference point for $\varepsilon_{1,1}(\nu,\xi)$ is $\xi=i\infty$, and that the path of integration Λ must be an $(i\nu\xi)$-progressive path linking $\xi=i\infty$ to $\xi=0$, or correspondingly $p=0$ to $p=(1+\tau^2)^{-1/2}$. (For a definition of a progressive path, see p. 222 of Chap. 6.)

Our choice of Λ is as follows. For large positive R let Γ_R be the path linking $\xi=iR$ to $\xi=0$, consisting of the union of a circular arc from $\xi=iR$ to $\xi=R$ with a real segment from $\xi=R$ to $\xi=0$. It is seen that Γ_R is an $(i\nu\xi)$-progressive path. The

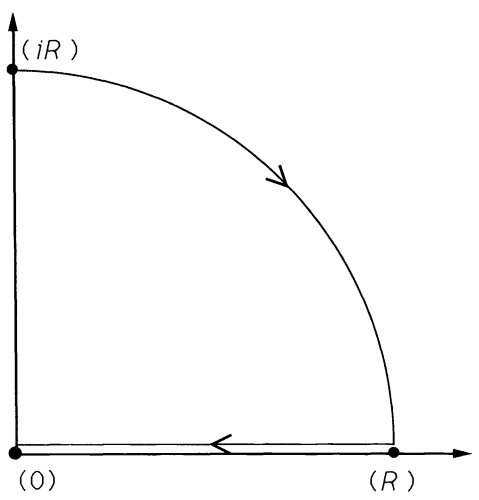


FIG. 2a. *Path Γ_R in ξ plane.*

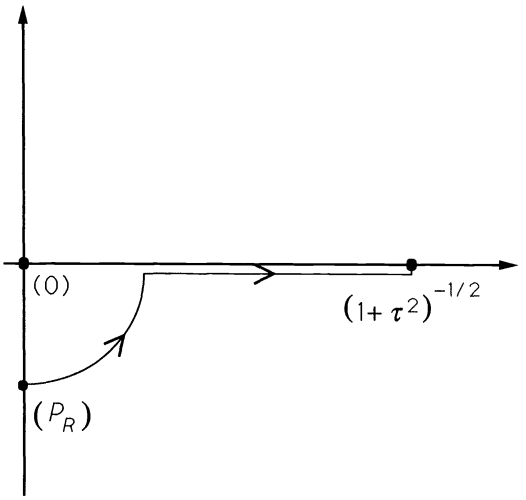


FIG. 2b. *Path γ_R in p plane.*

corresponding path in the p plane, γ_R, links $p=p_R$ to $p=(1+\tau^2)^{-1/2}$, where $p_R=-i/R+O(iR^{-2})$; see Fig. 2a, b. We take our variation path Λ to be the limit of Γ_R as $R\to\infty$. In the p plane the ξ-path Λ corresponds to the real segment $0\leqq p\leqq(1+\tau^2)^{-1/2}$; note that we can neglect the contribution to the variation from the vanishingly small arc near $p=0$.

Thus with our choice of Λ

$$\text{(A.3)}\qquad \mathscr{V}_\Lambda(U_1)=\int_0^{(1+\tau^2)^{-1/2}}\frac{|1-5p^2|}{8}\,dp=\frac{1}{6\sqrt{5}}+\frac{2-3\tau}{24(1+\tau^2)^{3/2}}=0.0091\cdots.$$

Now, since the RHS of (A.1) is a monotonically decreasing function of ν, it follows that it is bounded above by $1/\sqrt{2}$ if

$$\text{(A.4)}\qquad 2\mathscr{V}_\Lambda(U_1)/\nu<\nu_0,$$

where ν_0 is the root of the equation

$$\text{(A.5)}\qquad \nu_0\,e^{\nu_0}=1/\sqrt{2}\qquad(\nu_0=0.4506\cdots).$$

By symmetry $|\varepsilon_{1,2}(\nu,0)|$ is also bounded above by $1/\sqrt{2}$ if (A.4) holds, and so, in conclusion, we have shown that a sufficient condition for (3.23) to hold is for

$$\text{(A.6)}\qquad \nu>2\mathscr{V}_\Lambda(U_1)/\nu_0=0.4039\cdots.$$

Acknowledgments. I thank Professor F. W. J. Olver for a number of valuable suggestions, and the referee for some detailed and helpful comments. Also, I thank Dr. Matthew Yedlin and Dr. Norman Peterson for references on physical and geophysical applications.

REFERENCES

M. ABRAMOWITZ AND I. A. STEGUN (1965), *Handbook of Mathematical Functions*, Seventh edition, Dover, New York.

C. B. BALOGH (1967), *Asymptotic expansions of the modified Bessel function of the third kind of imaginary order*, SIAM J. Appl. Math., 15, pp. 1315-1323.

L. M. BREKHOVSKIKH (1960), *Waves in Layered Media*, Academic Press, London.

A. F. DE O. FALCÃO (1973), *Asymptotic expansions of functions related to imaginary order Bessel functions*, Z. Angew. Math. Mech., 58, pp. 133–134.

E. M. FERREIRA AND J. SESMA (1970), *Zeros of modified Hankel functions*, Numer. Math., 16, pp. 278–284.

A. GRAY, G. B. MATHEWS, AND T. M. MACROBERT (1952), *A Treatise on Bessel Functions and Their Applications to Physics*, Macmillan, London.

R. N. GUPTA (1965), *Reflection of plane waves from a linear transition layer in liquid media*, Geophys., 30, pp. 122–132.

P. W. HEMKER (1974), *The numerical solution of a singular perturbation problem in the domain exterior of a circle*, Mathematische Centrum, Amsterdam.

H. JEFFREYS (1962), *Asymptotic Approximations*, Oxford University Press, Oxford, London.

C. J. JOACHAIN (1975), *Quantum Collision Theory*, North-Holland, Amsterdam, New York.

V. I. KOGAN AND V. M. GALITSKY (1963), *Problems in Quantum Mechanics*, Prentice-Hall, Englewood Cliffs, NJ.

A. LAFORGIA (1986), *Inequalities and monotonicity results for zeros of modified Bessel functions of purely imaginary order*, Quart. Appl. Math., 44, pp. 91–96.

W. MAGNUS, F. OBERHETTINGER, AND R. P. SONI (1966), *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, Berlin, New York.

J. C. P. MILLER (1950), *On the choice of standard solutions for a homogeneous linear differential equation of the second order*, Quart. J. Mech. Appl. Math., 3, pp. 225–235.

F. W. J. OLVER (1974), *Asymptotics and Special Functions*, Academic Press, New York.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 1019-1030, July 1990
© 1990 Society for Industrial and Applied Mathematics
011

GENERAL ANALYTIC SOLUTION OF CERTAIN FUNCTIONAL EQUATIONS OF ADDITION TYPE*

M. BRUSCHI† AND F. CALOGERO†

Abstract. The general analytic solutions of the following functional equations are exhibited:

$$\alpha(x+y)/[\alpha(x)\alpha(y)]=1+\varphi(x)\varphi(y)\psi(x+y),$$

$$\beta(x+y)/[\beta(x)\beta(y)]=\gamma(x)+\gamma(y)+\chi(x+y).$$

These solutions are expressed in terms of Weierstrass elliptic functions; the special cases in which these reduce to elementary functions are also exhibited. Moreover, several remarkable formulae satisfied by Weierstrass elliptic functions are reported.

Key words. functional equations, special functions

AMS(MOS) subject classifications. 39B40, 33A25

1. Introduction. The main purpose of this paper is to report the general analytic solution of the following two functional equations:

$$\alpha(x+y)/[\alpha(x)\alpha(y)]=1+\varphi(x)\varphi(y)\psi(x+y), \tag{1.1}$$

$$\beta(x+y)/[\beta(x)\beta(y)]=\gamma(x)+\gamma(y)+\chi(x+y). \tag{1.2}$$

Actually the second of these equations is a limiting case of the first, as we will show in § 4; we prefer nevertheless to treat them separately.

Clearly these functional equations are susceptible to many reformulations, which we obtain by redefining the a priori unknown functions, namely $\alpha(z)$, $\varphi(z)$, and $\chi(z)$ in (1.1), and $\beta(z)$, $\gamma(z)$, and $\chi(z)$ in (1.2). For instance, other avatars of (1.1) read as follows:

$$\alpha(x+y)/[\alpha(x)\alpha(y)]=1+\psi(x+y)/[\omega(x)\omega(y)], \tag{1.1a}$$

$$\alpha(x)\alpha(y)/\alpha(x+y)=1+\Phi(x)\Phi(y)\Psi(x+y), \tag{1.1b}$$

$$\alpha(x)\alpha(y)/\alpha(x+y)=1-\Phi(x)\Phi(y)/\Omega(x+y), \tag{1.1c}$$

$$\alpha(x+y)\Omega(x+y)=\Phi(x)\Phi(y)\alpha(x+y)+\alpha(x)\alpha(y)\Omega(x+y), \tag{1.1d}$$

$$\alpha(x+y)-\alpha(x)\alpha(y)=\Phi(x)\Phi(y)\psi(x+y), \tag{1.1e}$$

$$\alpha(x+y)\omega(x)\omega(y)-\psi(x+y)\alpha(x)\alpha(y)=\alpha(x)\alpha(y)\omega(x)\omega(y), \tag{1.1f}$$

$$\ln[\alpha(x+y)-\alpha(x)\alpha(y)]=f(x)+f(y)+g(x+y), \tag{1.1g}$$

$$\ln[1-\alpha(x)\alpha(y)/\alpha(x+y)]=f(x)+f(y)+h(x+y), \tag{1.1h}$$

and other avatars of (1.2) read as follows:

$$\theta(x)\theta(y)/\theta(x+y)=\gamma(x)+\gamma(y)+\chi(x+y), \tag{1.2a}$$

$$b(x+y)-b(x)-b(y)=\ln[\gamma(x)+\gamma(y)+\chi(x+y)], \tag{1.2b}$$

$$\exp\{\beta(x+y)/[\beta(x)\beta(y)]\}=G(x)G(y)H(x+y), \tag{1.2c}$$

$$\exp[\theta(x)\theta(y)/\theta(x+y)]=G(x)G(y)H(x+y). \tag{1.2d}$$

* Received by the editors March 10, 1989; accepted for publication (in revised form) August 24, 1989. This research was partly supported by the Italian Ministry of Education.

† Istituto Nazionale di Fisica Nucleare, Sezione di Roma, Rome, Italy, and Dipartimento di Fisica, Università degli Studi di Roma "La Sapienza," Rome, Italy.

The key to these transformations reads as follows: $\omega = 1/\varphi$, $\Phi = \alpha\varphi$, $\Psi = -\psi/\alpha$, $\Omega = \alpha/\psi$, $f = \ln(\alpha\varphi)$, $g = \ln(\psi)$, $h = \ln(\psi/\alpha)$, $\theta = 1/\beta$, $b = \ln(\beta)$, $G = \exp(\gamma)$, $H = \exp(\chi)$. In the following we refer for definiteness to the canonical forms (1.1) and (1.2).

In § 2 we report the general analytic solutions of the functional equations (1.1) and (1.2). In § 3 we motivate our interest in these functional equations. In § 4 we prove our results. In § 5 we display several remarkable relations ("addition formulae" of various kinds) satisfied by the Weierstrass elliptic functions (whose definitions are collected in the Appendix, mainly to stake out our notation). Section 6 contains some concluding remarks.

2. Solutions. Before giving the solutions of the functional equations (1.1) and (1.2), let us mention some invariance properties of these equations.

It is plain that, if $\alpha(z)$, $\varphi(z)$, and $\psi(z)$ satisfy (1.1), so do

$$\begin{aligned} \tilde{\alpha}(z) &= \exp(bz)\alpha(az), \\ \tilde{\varphi}(z) &= A\exp(cz)\varphi(az), \\ \tilde{\psi}(z) &= A^{-2}\exp(-cz)\psi(az), \end{aligned} \tag{2.1}$$

with a, b, c, A arbitrary constants ($A \neq 0$), as well as

$$\begin{aligned} \tilde{\alpha}(z) &= 1/\alpha(z), \\ \tilde{\varphi}(z) &= \varphi(z)\alpha(z), \\ \tilde{\psi}(z) &= -\psi(z)/\alpha(z). \end{aligned} \tag{2.2}$$

Similarly, if $\beta(z)$, $\gamma(z)$, and $\chi(z)$ satisfy (1.2), so do

$$\begin{aligned} \tilde{\beta}(z) &= C\exp(bz)\beta(az), \\ \tilde{\gamma}(z) &= C^{-1}[\gamma(az) + Az + B], \\ \tilde{\chi}(z) &= C^{-1}[\chi(az) - Az - 2B], \end{aligned} \tag{2.3}$$

with a, b, A, B, and C arbitrary constants ($C \neq 0$).

The general analytic solution of the functional equation (1.1) reads as follows:

$$\alpha(z) = \exp(bz)\sigma(\mu)\sigma(az+\nu)/[\sigma(\nu)\sigma(az+\mu)], \tag{2.4a}$$

$$\varphi(z) = A\exp(cz)\sigma(az)/\sigma(az+\nu), \tag{2.4b}$$

$$\psi(z) = A^{-2}\exp(-cz)\sigma(\nu-\mu)\sigma(az+\mu+\nu)/[\sigma(\mu)\sigma(az+\mu)]. \tag{2.4c}$$

Here $\sigma(z) \equiv \sigma(z\,|\,\omega, \omega')$ is the Weierstrass σ-function (see the Appendix), and A, a, b, c, μ, ν, ω, and ω' are eight constants (arbitrary, except for the trivial restrictions needed to make good sense of the right-hand side (r.h.s.) of (2.4a-c)).

In § 4 we prove that functions (2.4) satisfy (1.1), and we show moreover that any analytic function $\alpha(z)$ satisfying (1.1) may depend on at most six free parameters. The fact that the expression (2.4a) of $\alpha(z)$ indeed contains six arbitrary parameters, namely a, b, μ, ν, ω, and ω', justifies our claim that formulae (2.4) provide the *general* analytic solution of the functional equation (1.1).

It may be easily verified that the solutions (2.4) are consistent with the transformations (2.1), (2.2), whose only effect is to cause a redefinition of some parameters.

For special choices of the parameters, (2.4) may be cast in simpler form. For instance the following expressions of $\alpha(z)$ in terms of Jacobi elliptic functions (see the Appendix) are all special cases of (2.4a):

$$\alpha(z) = \mathrm{sn}(\mu)/\mathrm{sn}(z+\mu), \tag{2.5a}$$

$$\alpha(z) = \mathrm{sn}(\mu)\,\mathrm{cn}(z+\mu)/[\mathrm{cn}(\mu)\,\mathrm{sn}(z+\mu)], \tag{2.5b}$$

$$\alpha(z) = \mathrm{sn}(\mu)\,\mathrm{dn}(z+\mu)/[\mathrm{dn}(\mu)\,\mathrm{sn}(z+\mu)], \tag{2.5c}$$

$$\alpha(z) = \mathrm{cn}(z+\mu-\omega_3)/\mathrm{cn}(\mu-\omega_3), \tag{2.5d}$$

$$\alpha(z) = \mathrm{dn}(z+\mu-\omega_3)/\mathrm{dn}(\mu-\omega_3), \tag{2.5e}$$

$$\alpha(z) = \mathrm{dn}(\mu-\omega_2)\,\mathrm{cn}(z+\mu+\omega_2)/[\mathrm{cn}(\mu-\omega_2)\,\mathrm{dn}(z+\mu+\omega_2)]. \tag{2.5f}$$

In these formulae μ is an arbitrary constant, and we use the standard notation for the Jacobi functions and their "periods" (see the Appendix).

More special (and perhaps more interesting) cases obtain when one or both periods of the elliptic functions diverge and they reduce to elementary functions (see the Appendix). The corresponding formulae for $\alpha(z)$, $\varphi(z)$, and $\psi(z)$ read as follows:

$$\alpha(z) = \exp\{[b+(\mu-\nu)a/3]z\}\sinh(\mu)\sinh(az+\nu)/[\sinh(\nu)\sinh(az+\mu)], \tag{2.6a}$$

$$\varphi(z) = A\exp[\nu^2/6+(c+\nu a/3)z]\sinh(az)/\sinh(az+\nu), \tag{2.6b}$$

$$\psi(z) = A^{-2}\exp[-\nu^2/3-(c+\nu a/3)z]\frac{\sinh(\nu-\mu)\sinh(az+\mu+\nu)}{\sinh(\mu)\sinh(az+\mu)}, \tag{2.6c}$$

$$\alpha(z) = \exp(bz)(\mu/\nu)(az+\nu)/(az+\mu), \tag{2.7a}$$

$$\varphi(z) = A\exp(cz)az/(az+\nu), \tag{2.7b}$$

$$\psi(z) = A^{-2}\exp(-cz)(\nu-\mu)(az+\mu+\nu)/[\mu(az+\mu)]. \tag{2.7c}$$

Note that the trivial solution $\alpha(z) = 1/C$, $\varphi(z) = A$, $\psi(z) = (C-1)/A^2$ with C and A arbitrary constants, obtains if, in (2.7), we set $b = c = 0$, $\mu = \delta$, $\nu = C\delta$, $\delta \to 0$.

The general analytic solution of the functional equation (1.2) reads as follows:

$$\beta(z) = C\exp(bz)\sigma(\mu)\sigma(az)/\sigma(az+\mu), \tag{2.8a}$$

$$\gamma(z) = C^{-1}[Az+B+\zeta(az)], \tag{2.8b}$$

$$\chi(z) = C^{-1}[-Az-2B+\zeta(\mu)-\zeta(az+\mu)]. \tag{2.8c}$$

Here $\sigma(z) \equiv \sigma(z|\omega,\omega')$ is the Weierstrass σ-function and $\zeta(z) = \zeta(z|\omega,\omega') = \sigma'(z)/\sigma(z)$ is the Weierstrass ζ-function (see the Appendix); a, b, A, B, C, μ, ω, and ω' are eight arbitrary constants ($C \neq 0$).

As mentioned in § 1 and shown in § 4, the functional equation (1.2) may be obtained by an appropriate limiting procedure from the functional equation (1.1); likewise, (2.8a-c) may be derived, by an appropriate limiting procedure, from (2.4a-c). But since the limiting procedure is not trivial, we have considered it worthwhile to exhibit separately the two functional equations (1.1) and (1.2), as well as their general solutions (2.4) and (2.8). We also report here the special cases of (2.8) analogous to the special cases of (2.4) displayed above (see (2.5)-(2.7)). These formulae read as follows:

$$\beta(z) = \mathrm{sn}(z), \tag{2.9a}$$

$$\beta(z) = \mathrm{sn}(z)/\mathrm{cn}(z), \tag{2.9b}$$

$$\beta(z) = \mathrm{sn}(z)/\mathrm{dn}(z), \tag{2.9c}$$

(2.10a) $$\beta(z) = C \exp[(b+\mu a/3)z] \sinh(\mu) \sinh(az)/\sinh(az+\mu),$$

(2.10b) $$\gamma(z) = C^{-1}[(A-a/3)z + B + \coth(az)],$$

(2.10c) $$\chi(z) = C^{-1}[-(A-a/3)z - 2B + \coth(\mu) - \coth(az+\mu)],$$

(2.11a) $$\beta(z) = C \exp(bz) \mu az/(az+\mu),$$

(2.11b) $$\gamma(z) = C^{-1}[Az + B + 1/(az)],$$

(2.11c) $$\chi(z) = C^{-1}[-Az - 2B + 1/\mu - 1/(az+\mu)].$$

Note that the trivial solution $\beta(z) = C$, with C an arbitrary constant, obtains if we set $b = 0$ in (2.11) and take the limit $a \to \infty$.

3. Motivation. Some years ago the (differential) functional equation

(3.1) $$\theta(x)\theta'(y) - \theta'(x)\theta(y) = \theta(x+y)[\varepsilon(x) - \varepsilon(y)]$$

was obtained and solved in the context of the study of a certain class of integrable dynamical systems [1]. Recently, in an analogous context, an analogous (differential) functional equation was obtained and solved [2]:

(3.2) $$\alpha(x)\alpha'(y) - \alpha'(x)\alpha(y) = [\alpha(x+y) - \alpha(x)\alpha(y)][\eta(x) - \eta(y)].$$

Clearly these two functional equations may be unified by considering the following functional equation:

(3.3) $$\tilde{\alpha}(x)\tilde{\alpha}'(y) - \tilde{\alpha}'(x)\tilde{\alpha}(y) = [\tilde{\alpha}(x+y) - c\tilde{\alpha}(x)\tilde{\alpha}(y)][\tilde{\eta}(x) - \tilde{\eta}(y)].$$

Indeed (up to notational changes) this equation yields (3.1) for $c = 0$ and (3.2) for $c = 1$. Moreover, provided $c \neq 0$, (3.3) coincides with (3.2) after the trivial rescalings

(3.4) $$\alpha(z) = c\tilde{\alpha}(z), \qquad \eta(z) = c\tilde{\eta}(z).$$

We now show, following [2], that the functional equation (3.3) may be integrated to yield (1.1) and (1.2). Let

(3.5) $$F(x, y) = c^{-1} \ln[1 - c\tilde{\alpha}(x)\tilde{\alpha}(y)/\tilde{\alpha}(x+y)].$$

It is then easily seen that (3.3) implies the first-order PDE

(3.6) $$F_x(x, y) - F_y(x, y) = \tilde{\eta}(x) - \tilde{\eta}(y),$$

whose general solution reads

(3.7) $$F(x, y) = H(x+y) + E(x) + E(y),$$

with $H(z)$ arbitrary and

(3.8) $$E(z) = \int^z dz' \, \tilde{\eta}(z').$$

Now note that (3.5) and (3.7) imply the relation

(3.9) $$1 - c\tilde{\alpha}(x)\tilde{\alpha}(y)/\tilde{\alpha}(x+y) = \exp\{c[H(x+y) + E(x) + E(y)]\}.$$

For $c \neq 0$ this equation coincides with (1.1) via the positions

(3.10) $$\alpha(z) = c\tilde{\alpha}(z), \quad \varphi(z) = \exp[cE(z)]/\alpha(z), \quad \psi(z) = \alpha(z) \exp[cH(z)].$$

And the treatment remains valid also in the limit $c \to 0$, in which case (3.9) yields (1.2) (up to notational changes; see § 4 for details).

4. Proofs. The validity of the "invariance properties" (2.1)-(2.3) is verified trivially.

Our first task is to prove that (2.4) satisfies (1.1). But (2.1) implies that, to prove this, it is sufficient to verify that (1.1) is satisfied by the following functions:

$$\alpha(z)=\sigma(\mu)\sigma(z+\nu)/[\sigma(\nu)\sigma(z+\mu)], \tag{4.1a}$$

$$\varphi(z)=\sigma(z)/\sigma(z+\nu), \tag{4.1b}$$

$$\psi(z)=\sigma(\nu-\mu)\sigma(z+\mu+\nu)/[\sigma(\mu)\sigma(z+\mu)] \tag{4.1c}$$

(corresponding to (2.4) with $b=c=0$ and $a=A=1$).

This has already been proved in [2], but in a somewhat cumbersome manner. A more straightforward proof may be based on the general "addition formula" (see § 5)

$$\begin{aligned}&\sigma(u+v_1)\sigma(u-v_1)\sigma(v_2+v_3)\sigma(v_2-v_3)\\ &\quad+\sigma(u+v_2)\sigma(u-v_2)\sigma(v_3+v_1)\sigma(v_3-v_1)\\ &\quad+\sigma(u+v_3)\sigma(u-v_3)\sigma(v_1+v_2)\sigma(v_1-v_2)=0.\end{aligned} \tag{4.2}$$

Indeed it is easily seen that the insertion of (4.1) into (1.1) yields precisely (4.2), with

$$u=(x+y)/2,\quad v_1=(x+y)/2+\nu,\quad v_2=(x-y)/2,\quad v_3=-(x+y)/2-\mu. \tag{4.3}$$ □

We now prove that any analytic solution $\alpha(z)$ of the functional equation (1.1) may contain at most six free parameters. It is actually expedient to base this proof on the differential functional equation (3.2), which is implied by (1.1), as shown in § 3. We set $y=\delta$ in (3.2), expand around $\delta=0$, and equate to zero the coefficients of δ^n, using the ansatzen

$$\alpha(\delta)=\alpha_0+\alpha_1\delta+\tfrac{1}{2}\alpha_2\delta^2+\tfrac{1}{6}\alpha_3\delta^3+o(\delta^4), \tag{4.4a}$$

$$\eta(\delta)=\eta_{-1}\delta^{-1}+\eta_0+\eta_1\delta+o(\delta^2), \tag{4.4b}$$

whose justification is implied a posteriori by the consistency of the following results. We thus get (for $n=-1, 0, 1, 2$)

$$\alpha_0=1, \tag{4.5a}$$

$$\eta_{-1}=1, \tag{4.5b}$$

$$\eta(z)=\eta_0+\tfrac{1}{2}[\alpha''(z)-2\alpha_1\alpha'(z)+\alpha_2\alpha(z)]/[\alpha'(z)-\alpha_1\alpha(z)], \tag{4.5c}$$

$$\begin{aligned}&2[\alpha'(z)-\alpha_1\alpha(z)]\alpha''(z)-3[\alpha''(z)-2\alpha_1\alpha'(z)]\alpha''(z)\\ &\quad+a_1[\alpha'(z)]^2+a_2\alpha'(z)\alpha(z)+a_3\alpha^2(z)=0,\end{aligned} \tag{4.5d}$$

with

$$a_1=6(2\eta_1-\alpha_2),\quad a_2=4(\alpha_3-6\alpha_1\eta_1),\quad a_3=3\alpha_2^2-4\alpha_1\alpha_3+12\alpha_1^2\eta_1. \tag{4.6}$$

Of course at each step we have used the findings from previous steps; note, incidentally, that (4.5c) provides an explicit definition of $\eta(z)$ in terms of $\alpha(z)$ (up to the parameter η_0, which remains completely arbitrary since it plays no role whatsoever; see (4.4b) and (3.2)).

This derivation implies that any analytic solution $\alpha(z)$ of the functional equation (1.1) must satisfy the constraint (4.5a) and the third-order (nonlinear) ODE (4.5d), which contains the four a priori undetermined parameters α_1, a_1, a_2, and a_3. Hence $\alpha(z)$ may depend at most on $6=4+3-1$ free parameters (the number 3 corresponds, of course, to the order of the ODE (4.5d), and -1 accounts for constraint (4.5a)). □

Analogously it can be shown that

$$\beta(z)=\sigma(\mu)\sigma(z)/\sigma(z+\mu), \tag{4.7a}$$

$$\gamma(z)=\zeta(z)\equiv\sigma'(z)/\sigma(z), \tag{4.7b}$$

$$\chi(z)=\zeta(\mu)-\zeta(z+\mu)\equiv\sigma'(\mu)/\sigma(\mu)-\sigma'(z+\mu)/\sigma(z+\mu) \tag{4.7c}$$

satisfy (1.2) (note that, via (2.3), this implies that (2.8) satisfies (1.2) as well). Indeed the insertion of these formulae in (1.2) yields the formula

$$\begin{aligned}\sigma(x+\mu)\sigma(y+\mu)\sigma(x+y)&=\sigma(\mu)\sigma(x+y+\mu)[\sigma(x)\sigma'(y)+\sigma'(x)\sigma(y)]\\&\quad+\sigma(x)\sigma(y)[\sigma(x+y+\mu)\sigma'(\mu)\\&\quad-\sigma'(x+y+\mu)\sigma(\mu)],\end{aligned} \tag{4.8}$$

whose validity is easily proved by setting in (4.2)

$$u=(x+y)/2+\mu,\quad v_1=(x+y)/2-\delta,\quad v_2=(x+y)/2,\quad v_3=(x+y)/2+\delta, \tag{4.9}$$

and then letting $\delta\to 0$.

But it is more interesting to prove that (1.2) is a limiting case of (1.1), and accordingly that the expressions (4.7) are a limiting case of (4.1). Indeed, setting

$$\alpha(z)=\delta^{-1}\beta(z),\quad \varphi(z)=1-\delta\gamma(z),\quad \psi(z)=-1+\delta\chi(z) \tag{4.10}$$

in the following equation:

$$\ln\{1-\alpha(x+y)/[\alpha(x)\alpha(y)]\}=\ln[\varphi(x)]+\ln[\varphi(y)]+\ln[-\psi(x+y)] \tag{4.11}$$

(which is clearly equivalent to (1.1)), and taking the $\delta\to 0$ limit under the assumption that in this limit the functions $\beta(z)$, $\gamma(z)$, and $\chi(z)$ remain finite, we find that (1.2) evidently obtains. On the other hand, the assumption about the finiteness of $\beta(z)$, $\gamma(z)$, and $\chi(z)$ is verified using the explicit expression (4.1) of $\alpha(z)$, $\varphi(z)$, and $\psi(z)$ with the position

$$\nu=\delta; \tag{4.12}$$

indeed, using (A.3)-(A.6) below, it is easily seen that (4.1) and (4.10) with (4.12) yield, in the $\delta\to 0$ limit, precisely (4.7). □

5. Addition formulae for Weierstrass elliptic functions. In this section we report some addition formulae for Weierstrass elliptic functions that, in spite of their remarkable neatness and generality, cannot be found in the standard compilations [3]-[5].

Foremost among these relations is the beautiful addition formula (4.2), which for completeness we report here in two equivalent forms:

$$\begin{aligned}&\sigma(u+v_1)\sigma(u-v_1)\sigma(v_2+v_3)\sigma(v_2-v_3)\\&\quad+\sigma(u+v_2)\sigma(u-v_2)\sigma(v_3+v_1)\sigma(v_3-v_1)\\&\quad+\sigma(u+v_3)\sigma(u-v_3)\sigma(v_1+v_2)\sigma(v_1-v_2)=0,\end{aligned} \tag{5.1a}$$

$$\begin{aligned}&\sigma(x+y)\sigma(y+z)\sigma(z+x)\sigma(2w)\\&\quad=\sigma(x+w)\sigma(y+w)\sigma(z+w)\sigma(x+y+z-w)\\&\quad-\sigma(x-w)\sigma(y-w)\sigma(z-w)\sigma(x+y+z+w),\end{aligned} \tag{5.1b}$$

related by the change of variables

$$u+v_1=x+y+z-w,\quad u-v_1=z+w,\quad v_2+v_3=x+w,\quad v_2-v_3=-(y+w). \tag{5.2}$$

Let us emphasize that this "addition formula" features four free parameters (in addition, of course, to the two "periods" of the Weierstrass σ-functions; see the Appendix). Formula (5.1a) is not new, (see, for instance, p. 389 of [6]); a straightforward way to prove it is by equating to zero the sum of the residues of the elliptic function

$$F(z)=\prod_{k=1}^{3}[\sigma(z-z_k)/\sigma(z-p_k)], \tag{5.3}$$

with the zeros z_k and poles p_k restricted by the condition

$$\sum_{k=1}^{3}(z_k-p_k)=0, \tag{5.4}$$

which is instrumental to guaranteeing that $F(z)$, as defined by (5.3), is indeed an elliptic function (hence that its residues within a fundamental parallelogram add up to zero). It is then easy to obtain (5.1a) with

$$u+v_1=z_2-z_3,\quad u-v_1=z_1-p_1,\quad v_2+v_3=z_1-p_2,\quad v_2-v_3=z_1-p_3. \tag{5.5}$$

Since the addition formula (5.1) features four free parameters, it is easy to obtain from it, merely by reduction, myriad addition formulae with three or two arguments, including all the "classical" formulae that can be found in the standard compilations, and many others that are less advertised. For instance, setting $u=0$ in (5.1a) we get

$$\begin{aligned}&\sigma^2(v_1)\sigma(v_2+v_3)\sigma(v_2-v_3)+\sigma^2(v_2)\sigma(v_3+v_1)\sigma(v_3-v_1)\\&\quad+\sigma^2(v_3)\sigma(v_1+v_2)\sigma(v_1-v_2)=0,\end{aligned} \tag{5.6}$$

and setting $v_1=v_2+v_3$ in this formula yields (using (A.12))

$$\sigma(2v_1-v_2)\sigma^3(v_2)-\sigma(2v_2-v_1)\sigma^3(v_1)=\sigma(v_1-v_2)\sigma^3(v_1-v_2). \tag{5.7}$$

More generally, setting $u=\delta$ in (5.1a), expanding in δ, and equating the coefficients of δ^n obtains, in addition to (5.6) (which corresponds, of course, to $n=0$), the formula

$$\begin{aligned}&\mathcal{P}(v_1)\sigma^2(v_1)\sigma(v_2+v_3)\sigma(v_2-v_3)\\&\quad+\mathcal{P}(v_2)\sigma^2(v_2)\sigma(v_3+v_1)\sigma(v_3-v_1)\\&\quad+\mathcal{P}(v_3)\sigma^2(v_3)\sigma(v_1+v_2)\sigma(v_1-v_2)=0,\end{aligned} \tag{5.8}$$

which corresponds to $n=2$ ($n=1$ yields merely a trivial identity). To obtain this formula we have, of course, used the definition (A.4) of the Weierstrass $\mathcal{P}$-function. Note that, for $v_3=0$, (5.8) yields, using (A.6), (A.7), and (A.12), the standard addition formula (see (A.13))

$$\sigma(v_1+v_2)\sigma(v_1-v_2)=\sigma^2(v_1)\sigma^2(v_2)[\mathcal{P}(v_2)-\mathcal{P}(v_1)]. \tag{5.9}$$

On the other hand, taking the logarithmic derivative of (5.6) with respect to v_1 and using (A.3) obtains the formula

$$\begin{aligned}&\sigma(v_1+v_2)\sigma(v_1-v_2)\sigma^2(v_3)[2\zeta(v_1)-\zeta(v_1+v_2)-\zeta(v_1-v_2)]\\&\quad=\sigma(v_1+v_3)\sigma(v_1-v_3)\sigma^2(v_2)[2\zeta(v_1)-\zeta(v_1+v_3)-\zeta(v_1-v_3)],\end{aligned} \tag{5.10}$$

which, setting $v_3=v_1$, yields (using (A.6) and (A.7)),

$$\zeta(v_1+v_2)+\zeta(v_1-v_2)-2\zeta(v_1)=\sigma(2v_1)\sigma^2(v_2)/[\sigma^2(v_1)\sigma(v_1+v_2)\sigma(v_1-v_2)]. \tag{5.11}$$

This formula yields, via the duplication formula (see (A.17)),

$$\sigma(2z)=-\sigma^4(z)\mathcal{P}'(z), \tag{5.12}$$

the well-known relation (A.15),

$$\zeta(v_1+v_2)+\zeta(v_1-v_2)-2\zeta(v_1)=\mathcal{P}'(v_1)/[(\mathcal{P}(v_1)-\mathcal{P}(v_2)]. \tag{5.13}$$

Note, incidentally, that the duplication formula (5.12) may itself be derived, since (5.13) may be obtained directly from (5.9) (taking the logarithmic derivative with respect to v_1), and clearly (5.13) with (5.11) yields (5.12).

On the other hand, differentiating (5.6) with respect to v_1, we obtain the relation

(5.14)

$$\begin{aligned}&2\sigma(v_1)\sigma'(v_1)\sigma(v_2+v_3)\sigma(v_2-v_3)\\&\quad=\sigma^2(v_2)[\sigma(v_1+v_3)\sigma'(v_1-v_3)+\sigma(v_1-v_3)\sigma'(v_1+v_3)]-\sigma^2(v_3)\\&\qquad\cdot[\sigma(v_1+v_2)\sigma'(v_1-v_2)+\sigma(v_1-v_2)\sigma'(v_1+v_2)],\end{aligned} \tag{5.14}$$

and this, via (5.9), yields the formula

$$\begin{aligned}2\zeta(v_1)[\mathcal{P}(v_2)-\mathcal{P}(v_3)]&=[\zeta(v_1+v_2)+\zeta(v_1-v_2)][\mathcal{P}(v_2)-\mathcal{P}(v_1)]\\&\quad-[\zeta(v_1+v_3)+\zeta(v_1-v_3)][\mathcal{P}(v_3)-\mathcal{P}(v_1)].\end{aligned} \tag{5.15}$$

Let us also report some neat relations that are more conveniently obtained from (5.1b). In the limit $w\to 0$ this yields (via (A.3) and (A.6))

$$\begin{aligned}&\zeta(x)+\zeta(y)+\zeta(z)-\zeta(x+y+z)\\&\quad=\sigma(x+y)\sigma(y+z)\sigma(z+x)/[\sigma(x)\sigma(y)\sigma(z)\sigma(x+y+z)],\end{aligned} \tag{5.16}$$

which is essentially a more elegant version of (4.8); and in the limit $z\to 0$ this yields the standard formula (see (A.14))

$$\zeta(x+y)-\zeta(x)-\zeta(y)=\tfrac{1}{2}[\mathcal{P}'(x)-\mathcal{P}'(y)]/[\mathcal{P}(x)-\mathcal{P}(y)] \tag{5.17}$$

(which may also be easily derived from (5.13)).

On the other hand by setting $z=0$ in (5.1b) we obtain the relation

$$\begin{aligned}\sigma(x)\sigma(y)\sigma(x+y)\sigma(2w)/\sigma(w)&=\sigma(x+w)\sigma(y+w)\sigma(x+y-w)\\&\quad+\sigma(x-w)\sigma(y-w)\sigma(x+y+w),\end{aligned} \tag{5.18}$$

while differentiating (5.16) with respect to z yields (via (A.3)-(A.4), and again (5.16), and with the change of variables $x+y+z=u_1$, $z=u_2$, $y+z=-u_3$) the remarkably neat formula

$$\begin{aligned}\mathcal{P}(u_1)-\mathcal{P}(u_2)&=[\zeta(u_1)+\zeta(u_2)+\zeta(u_3)-\zeta(u_1+u_2+u_3)]\\&\quad\cdot[\zeta(u_1)-\zeta(u_1+u_3)-\zeta(u_2)+\zeta(u_2+u_3)].\end{aligned} \tag{5.19}$$

Note that the variable u_3 appears only on the right-hand side. Setting $u_3=\delta$ and expanding (5.19) around $\delta=0$, the coefficient of δ yields (5.9) while the coefficient of δ^2 yields (5.13).

6. Conclusion. It is in our opinion remarkable that the general analytic solutions of the functional equations (1.1) and (1.2), each featuring three a priori unknown functions, may be explicitly obtained.

These findings imply the possibility of obtaining solutions to more general functional equations. It is, for instance, clear from the results reported in § 2 (see, in particular, (2.4)) that the functional equation

$$\alpha(x+y)/[\alpha(x)\alpha(y)]=\prod_{n=1}^{N}[1+\varphi_n(x)\varphi_n(y)\psi_n(x+y)], \tag{6.1}$$

which features $2N+1$ a priori unknown functions and reduces to (1.1) for $N=1$, admits the solution

$$\alpha(z)=\exp(bz)\prod_{n=1}^{N}\{\sigma_n(\mu_n)\sigma_n(a_nz+\nu_n)/[\sigma_n(\nu_n)\sigma_n(a_nz+\mu_n)]\}, \tag{6.2a}$$

$$\varphi_n(z)=A_n\exp(c_nz)\sigma_n(a_nz)/\sigma_n(a_nz+\nu_n), \tag{6.2b}$$

$$\varphi_n(z)=A_n^{-2}\exp(-c_nz)\sigma_n(\nu_n-\mu_n)\sigma_n(a_nz+\mu_n+\nu_n)/[\sigma_n(\mu_n)\sigma_n(a_nz+\mu_n)]. \tag{6.2c}$$

Here (and below) we use for the Weierstrass σ-functions the abbreviated notation $\sigma_n(z)\equiv\sigma(z\,|\,\omega_n,\omega_n')$. Note that the solution (6.2a-c) contains $7N+1$ free parameters (namely A_n, a_n, b, c_n, μ_n, ν_n, ω_n, and ω_n', with $A_n\neq 0$), $5N+1$ of which enter into the expression (6.2a) of $\alpha(z)$.

It is likewise plain (see (2.8)) that the functional equation

$$\beta(x+y)/[\beta(x)\beta(y)]=\prod_{n=1}^{N}[\gamma_n(x)+\gamma_n(y)+\chi_n(x+y)], \tag{6.3}$$

which features $2N+1$ a priori unknown functions and reduces to (1.2) for $N=1$, admits the solution

$$\beta(z)=\exp(bz)\prod_{n=1}^{N}[C_n\sigma_n(\mu_n)\sigma_n(a_nz)/\sigma_n(a_nz+\mu_n)], \tag{6.4a}$$

$$\gamma_n(z)=C_n^{-1}[A_nz+B_n+\zeta_n(a_nz)], \tag{6.4b}$$

$$\chi_n(z)=C_n^{-1}[-A_nz-2B_n+\zeta_n(\mu_n)-\zeta_n(a_nz+\mu_n)]. \tag{6.4c}$$

Here, of course, $\zeta_n(z)\equiv\zeta(z\,|\,\omega_n,\omega_n')$. Note also that this solution contains $7N+1$ free parameters (namely a_n, b, A_n, B_n, C_n, μ_n, ω_n, and ω_n'), $5N+1$ of which enter in the definition (6.4a) of $\beta(z)$.

A question that remains open for the moment is whether (6.2a-c), respectively, (6.4a-c), are the *general* analytic solutions of (6.1), respectively, (6.3).

Of course, many other functional equations, whose solutions can be easily found from the solutions of (1.1) and (1.2), may be manufactured combining (1.1) and (1.2) and/or their avatars (see, for instance, (1.1a-h) and (1.2a-d)).

These functional equations, together with their solutions, provide moreover a convenient tool for uncovering additional relations, satisfied by Weierstrass elliptic functions, that are generally all consequences of (5.1) but might be quite difficult to discover by direct computation. It is, for instance, plain (from the relation $\theta=1/\beta$ and from (4.7) with μ replaced by ν) that the functional equation (1.2a) admits the solution

$$\tilde{\theta}(z)=\sigma(z+\nu)/[\sigma(\nu)\sigma(z)], \tag{6.5a}$$

$$\tilde{\gamma}(z)=\zeta(z), \tag{6.5b}$$

$$\tilde{\chi}(z)=\zeta(\nu)-\zeta(z+\nu). \tag{6.5c}$$

On the other hand, it is evident that the relation

$$\alpha(z)=\beta(z)\tilde{\theta}(z) \tag{6.6}$$

holds, with $\alpha(z)$ defined by (4.1a), $\beta(z)$ defined by (4.7a), and $\tilde{\theta}(z)$ defined by (6.5a). Hence (1.1), (1.2), and (1.2a) imply the relation

$$[\gamma(x)+\gamma(y)+\chi(x+y)]/[\tilde{\gamma}(x)+\tilde{\gamma}(y)+\tilde{\chi}(x+y)]=1+\varphi(x)\varphi(y)\psi(x+y), \tag{6.7}$$

with φ, ψ, γ, χ, $\tilde{\gamma}$, respectively, $\tilde{\chi}$, defined by (4.1b, c), (4.7b, c), and (6.5b), respectively, (6.5c), namely the neat formula

$$\begin{aligned}(6.8)\quad &[\zeta(x)+\zeta(y)+\zeta(\mu)-\zeta(x+y+\mu)]/[\zeta(x)+\zeta(y)+\zeta(\nu)-\zeta(x+y+\nu)]\\ &\quad=1+\sigma(x)\sigma(y)\sigma(x+y+\mu)\sigma(\nu-\mu)/[\sigma(x+\nu)\sigma(y+\nu)\sigma(x+y+\mu)\sigma(\mu)].\end{aligned}$$

This formula, which contains four free parameters (in addition to ω and ω'), may also be obtained using (5.1a) *and* (5.1b).

Let us finally mention that a natural question suggested by the main findings of this paper concerns the solvability of the simplest functional equation that encompasses (1.1) and (1.2) and involves four a priori unknown functions, namely

$$(6.9)\qquad \alpha(x+y)/[\alpha(x)\alpha(y)]=\gamma(x)+\gamma(y)+\varphi(x)\varphi(y)\psi(x+y).$$

Appendix. For the sake of completeness and to standardize the notation, we report in this Appendix the relevant formulae for the Weierstrass functions $\sigma(z)$, $\zeta(z)$, $\mathcal{P}(z)$.

Definitions.

$$(A.1)\qquad w\equiv w_{m,n}\equiv 2m\omega+2n\omega',$$

$$(A.2)\qquad \sigma(z)=\sigma(z\,|\,\omega,\omega')=z\prod{}'\{(1-z/w)\exp[z/w+\tfrac{1}{2}(z/w)^2]\},$$

$$(A.3)\qquad \zeta(z)\equiv\zeta(z\,|\,\omega,\omega')=\sigma'(z)/\sigma(z),$$

$$(A.4)\qquad \mathcal{P}(z)\equiv\mathcal{P}(z\,|\,\omega,\omega')=-\zeta'=[\sigma'^2(z)-\sigma(z)\sigma''(z)]/\sigma^2(z).$$

Here and below, a prime appended to a function denotes differentiation, while $\prod'$ ($\sum'$) denotes the product (the sum) taken over all (positive and negative) integers m, n with the exception of $m=n=0$.

Laurent series.

$$(A.5)\qquad g_2=60\sum{}' w^{-4},\qquad g_3=140\sum{}' w^{-6},$$

$$(A.6)\qquad \sigma(z)=\sum_{m=n=0}^{\infty} a_{m,n}(g_2/2)^m(2g_3)^n z^{(4m+6n+1)}/(4m+6n+1)!,$$

where

$$(A.6a)\qquad a_{0,0}=1,\qquad a_{m,n}=0\quad\text{if } m<0\quad\text{or}\quad n<0,$$

$$\begin{aligned}(A.6b)\quad a_{m,n}&=(3m+1)a_{m+1,n-1}+\tfrac{16}{3}(n+1)a_{m-2,n+1}\\ &\quad-\tfrac{1}{3}(3m+3n-1)(4m+6n-1)a_{m-1,n},\end{aligned}$$

$$(A.7)\qquad \zeta(z)=z^{-1}-\sum_{k=2}^{\infty} c_k z^{2k-1}/(2k-1),$$

where

$$(A.7a)\qquad c_2=g_2/20,\qquad c_3=g_3/28,$$

$$(A.7b)\qquad c_k=3/[(2k+1)(k-3)]\sum_{m=2}^{k-2} c_m c_{k-m},\qquad k\geqq 4,$$

$$(A.8)\qquad \mathcal{P}(z)=z^{-2}+\sum_{k=2}^{\infty} c_k z^{2(k-1)}.$$

Period relations.

$$\sigma(z+2m\omega+2n\omega') = (-1)^{(m+n+mn)}\sigma(z) \times \exp[(z+m\omega+n\omega')(2m\zeta(\omega)+2n\zeta(\omega'))], \tag{A.9}$$

$$\zeta(z+2m\omega+2n\omega') = \zeta(z)+2m\zeta(\omega)+2n\zeta(\omega'), \tag{A.10}$$

$$\mathcal{P}(z+2m\omega+2n\omega') = \mathcal{P}(z). \tag{A.11}$$

Functional relations and properties.

$$\sigma(z) = -\sigma(-z), \quad \zeta(z) = -\zeta(-z), \quad \mathcal{P}(z) = \mathcal{P}(-z), \tag{A.12}$$

$$\sigma(z_1+z_2)\sigma(z_1-z_2) = \sigma^2(z_1)\sigma^2(z_2)[\mathcal{P}(z_2)-\mathcal{P}(z_1)], \tag{A.13}$$

$$\zeta(z_1+z_2)+\zeta(z_1-z_2)-2\zeta(z_1) = \mathcal{P}'(z_1)/[\mathcal{P}(z_1)-\mathcal{P}(z_2)], \tag{A.14}$$

$$\zeta(z_1+z_2)-\zeta(z_1)-\zeta(z_2) = \tfrac{1}{2}[\mathcal{P}'(z_1)-\mathcal{P}'(z_2)]/[\mathcal{P}(z_1)-\mathcal{P}(z_2)], \tag{A.15}$$

$$\mathcal{P}(z_1+z_2)+\mathcal{P}(z_1)+\mathcal{P}(z_2) = \tfrac{1}{4}[\mathcal{P}'(z_1)-\mathcal{P}'(z_2)]^2/[\mathcal{P}(z_1)-\mathcal{P}(z_2)]^2, \tag{A.16}$$

$$\sigma(2z) = -\mathcal{P}'(z)\sigma^4(z), \tag{A.17}$$

$$\zeta(2z) = 2\zeta(z)+\tfrac{1}{2}\mathcal{P}''(z)/\mathcal{P}(z), \tag{A.18}$$

$$\mathcal{P}(2z) = -2\mathcal{P}(z)+\tfrac{1}{4}[\mathcal{P}''(z)/P'(z)]^2. \tag{A.19}$$

Connection with the Jacobian elliptic functions.

$$\omega_1 = \omega, \quad \omega_2 = -(\omega+\omega'), \quad \omega_3 = \omega', \tag{A.20a}$$

$$\eta_k = \zeta(\omega_k), \qquad k = 1, 2, 3, \tag{A.20b}$$

$$e_k = \mathcal{P}(\omega_k), \qquad k = 1, 2, 3, \tag{A.20c}$$

$$\sigma_k(z) = \exp(-\eta_k z)\sigma(z+\omega_k)/\sigma(\omega_k), \qquad k = 1, 2, 3, \tag{A.20d}$$

$$u = (e_1-e_3)^{1/2}z, \qquad m^2 = (e_2-e_3)/(e_1-e_3), \tag{A.21}$$

$$\operatorname{sn}(u|m) = (e_1-e_3)^{1/2}\sigma(z)/\sigma_3(z), \tag{A.22a}$$

$$\operatorname{cn}(u|m) = \sigma_1(z)/\sigma_3(z), \tag{A.22b}$$

$$\operatorname{dn}(u|m) = \sigma_2(z)/\sigma_3(z). \tag{A.22c}$$

Degenerate cases.

$$e_1 = e_2 = a, \quad e_3 = -2a, \quad \omega = \infty, \quad \omega' = (12a)^{-1/2}\pi i, \tag{A.23a}$$

$$\sigma(z) = (3a)^{-1/2}\sinh[(3a)^{1/2}z]\exp[-az^2/2], \tag{A.23b}$$

$$\zeta(z) = -az+(3a)^{1/2}\coth[(3a)^{1/2}z], \tag{A.23c}$$

$$\mathcal{P}(z) = a+3a\{\sinh[(3a)^{1/2}z]\}^{-2}, \tag{A.23d}$$

$$e_1 = e_2 = e_3 = 0, \qquad \omega = -i\omega' = \infty, \tag{A.24a}$$

$$\sigma(z) = z, \tag{A.24b}$$

$$\zeta(z) = 1/z, \tag{A.24c}$$

$$\mathcal{P}(z) = 1/z^2. \tag{A.24d}$$

Acknowledgments. After obtaining (5.1) in a cumbersome manner, we learned from G. Chudnovsky, via R. Askey, that this formula could be found in the literature. We thank both these friends for providing this information, and moreover we thank R. Askey for carefully reading this paper and suggesting some improvements.

REFERENCES

[1] F. CALOGERO, *Exactly solvable one-dimensional many-body problems*, Lett. Nuovo Cimento (2), 13 (1975), pp. 411–416.

[2] M. BRUSCHI AND F. CALOGERO, *The Lax representation for an integrable class of relativistic dynamical systems*, Comm. Math. Phys., 109 (1987), pp. 3–23.

[3] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Applied Mathematics Series 55, National Bureau of Standards, Washington, DC, 1964.

[4] I. S. GRADSHTEYN AND I. M. RYZHIK, *Table of Integrals, Series and Products*, Academic Press, New York, 1965.

[5] A. ERDELYI, ED., *Higher Transcendental Functions*, Vol. II, McGraw-Hill, New York, 1953.

[6] H. HANCOCK, *Lectures on the Theory of Elliptic Functions*, John Wiley, New York, 1910.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 1031-1050, July 1990
© 1990 Society for Industrial and Applied Mathematics
012

THE RECOVERY OF ORTHOGONAL POLYNOMIALS FROM A SUM OF SQUARES*

B. F. LOGAN†

Abstract. It is shown that a positive polynomial of degree $2n$ has a unique representation as the sum of squares of polynomials of degrees 0 through n if the polynomials are real-valued and orthonormal with respect to some positive measure. The polynomials may be found by solving an electrostatic equilibrium problem.

Key words. orthogonal polynomials, positive polynomial representations, osculatory interpolation with a sum of squares, second-order differential equations

AMS(MOS) subject classifications. primary 42C05; secondary 31A35, 34A05

1. Introduction. Consider a set $\{p_k\}_0^n$ of polynomials of degree k, $k=0, 1, \cdots, n$, having leading coefficients 1 and orthogonal with respect to a positive measure $d\alpha$ on the real line of total mass μ_0,

$$\int_{-\infty}^{\infty} p_j(x)p_k(x)\,d\alpha(x) = \delta_{jk}h_k \qquad (h_0=\mu_0). \tag{1.1}$$

The subspace of $L^2(d\alpha)$ spanned by polynomials of degree n has the reproducing kernel

$$K_n(x, t) = \sum_{k=0}^{n} \frac{p_k(x)p_k(t)}{h_k}. \tag{1.2}$$

The problem considered here is that of recovering the orthogonal polynomials $\{p_k\}_0^n$ from their "trace" (of order n), which is the name we give to the function

$$K_n(x, x) = \sum_{k=0}^{n} \frac{\{p_k(x)\}^2}{h_k} \qquad (-\infty<x<\infty), \tag{1.3}$$

where n is considered fixed in the problem. No matter how the polynomials are normalized, the reproducing kernel is unique. It does depend in a simple way on scaling of the measure $d\alpha(x)$. Thus by leaving the total mass μ_0 unspecified we have the desirable result that $\lambda K_n(x, x)$, $\lambda>0$, is a trace of $\{p_k\}_0^n$ whenever $K_n(x, x)$ is. In other words, if the trace determines the polynomials (leading coefficients = 1), it also determines μ_0. Recall that the trace has an important meaning which derives from the extremal property of the reproducing kernel.

Let $\pi_n(t)$ be a polynomial of degree n having norm 1 in $L^2(d\alpha(t))$. Then how large can $|\pi_n(t)|$ be at any given real point, say $t=x$? We have

$$\pi_n(x) = \int_{-\infty}^{\infty} K_n(x, t)\pi_n(t)\,d\alpha(t).$$

Then by Schwarz's inequality,

$$|\pi_n(x)|^2 \leqq \int_{-\infty}^{\infty} K_n^2(x, t)\,d\alpha(t) = K_n(x, x), \tag{1.4}$$

with equality possible if and only if

$$\pi_n(t) = c_n(x)K_n(x, t),$$

* Received by the editors May 22, 1989; accepted for publication August 28, 1989.
† AT&T Bell Laboratories, Murray Hill, New Jersey 07974.

where

$$|c_n(x)|^2 = \{K_n(x, x)\}^{-1}.$$

Now there are various discrete measures $d\alpha_n(t)$ consisting of point masses at certain real points $\{t_k\}$ and having the same moments through order $2n$ as $d\alpha(t)$ so that the quadrature formula,

$$\int_{-\infty}^{\infty} \pi_{2n}(t)\, d\alpha(t) = \sum_k m_n(t_k)\pi_{2n}(t_k),$$

is valid for any polynomial π_{2n} of degree $2n$. The question here is, how large can we make the mass $m_n(t_k)$ at a given real point, say $t_k = x$?

Again suppose that $\pi_n(t)$ is a polynomial of degree n having norm 1 in $L^2(d\alpha(t))$. Then, since $|\pi_n(t)|^2$ is a polynomial of degree $2n$ in t,

$$1 = \int_{-\infty}^{\infty} |\pi_n(t)|^2\, d\alpha_n(t) = \sum_k m_n(t_k)|\pi_n(t_k)|^2 \geqq m_n(x)|\pi_n(x)|^2.$$

Here we take the largest possible value for $|\pi_n(x)|^2$ as given by (1.4) to obtain

$$m_n(x) \leqq \rho_n(x) = \{K_n(x, x)\}^{-1}. \tag{1.5}$$

Equality can hold here if, and only if, the t_k are the zeros of $(t-x)K_n(x, t)$. Thus the reciprocal of the trace represents the most mass that any positive measure $d\alpha_n(t)$ may have at the point $t = x$, where the moments through order $2n$ are required to agree with those of $d\alpha(t)$. With this characterization of the trace, we would suspect that it determines the moments through order $2n$ and hence the polynomials $\{p_k\}_0^n$, as well as the norm of p_n. The polynomials, of course, are determined by the moments through order $2n-1$, leaving the norm of p_n undetermined. We certainly can read off the norm of p_n from the leading coefficient of the trace,

$$K_n(x, x) = \frac{x^{2n}}{h_n} + \cdots. \tag{1.6}$$

It should be noted that the analytic continuation $K_n(z, z)$ is not the usual quantity of interest, $K_n(z, \bar{z})$, which appears in the extension of (1.4) to complex variables.

The orthogonal polynomials, having leading coefficient 1, satisfy a recurrence relation of the form

$$p_{k+1}(x) = (x - b_k)p_k(x) - a_k p_{k-1}(x), \tag{1.7}$$

where

$$-\infty < b_k < \infty, \qquad a_k > 0.$$

We may suppose that this relation is satisfied for all positive k. In particular, it is convenient to introduce $p_{n+1}(x)$ by letting b_n be a free parameter not determined by $K_n(x, x)$. Note that if the trace does in fact determine the moments through order $2n$, then it determines a_n, for if both sides of (1.7) are multiplied by x^{k-1} and then integrated with respect to $d\alpha(x)$ there results the relation

$$0 = h_k - a_k h_{k-1},$$

and hence

$$h_k = \mu_0 \prod_{j=1}^{k} a_j. \tag{1.8}$$

If we were given p_n and p_{n-1}, then the recursion relation (1.7) could be worked backwards to determine the lower-order polynomials as well as the recurrence coefficients. So we may ask simply whether or not the trace determines $p_n(x)$ and $p_{n-1}(x)$.

In the problem here, where moments only through order $2n$ enter, the introduction of the undetermined orthogonal polynomial p_{n+1} may be regarded merely as a device for representing the reproducing kernel,

$$K_n(x, t) = \frac{1}{h_n} \frac{p_{n+1}(x)p_n(t) - p_{n+1}(t)p_n(x)}{x - t}, \tag{1.9}$$

for no matter what b_n is, substitution of

$$p_{n+1}(x) = (x - b_n)p_n(x) - a_n p_{n-1}(x) \tag{1.10}$$

in (1.9) gives, in view of (1.7) and (1.8),

$$K_n(x, t) = K_{n-1}(x, t) + \frac{p_n(x)p_n(t)}{h_n}, \tag{1.11}$$

and shows, by reduction to the case $n = 0$, that (1.9) is valid. Thus we may as well ask whether or not the trace $K_n(x, x)$, aside from obviously determining h_n, determines $p_n(x)$ and hence $K_{n-1}(x, x)$.

From (1.9) we have

$$K_n(x, x) = \frac{W_n(x)}{h_n}, \tag{1.12}$$

where

$$W_n(x) = p'_{n+1}(x)p_n(x) - p_{n+1}(x)p'_n(x), \tag{1.13}$$

is the Wronskian of p_{n+1} and p_n, being a positive polynomial of degree $2n$ with leading coefficient 1. Now the only condition that two polynomials $p_{n+1}(x)$ and $p_n(x)$ need satisfy in order to be consecutive orthogonal polynomials is that their zeros be real and interlaced [1], for two such polynomials, with leading coefficients 1, determine the recursion coefficients b_k, $0 \leqq k \leqq n$, and a_k, $1 \leqq k \leqq n$, in (1.7), the interlacing of the zeros implying the positivity of the a_k. Hence the Wronskian of two such polynomials is positive, since it may be identified as the trace (with $h_n = 1$) of some system $\{p_k\}_0^n$ of orthogonal polynomials, all of which may be found by working the recurrence relation backwards, determining the recurrence coefficients in the process. So we may equivalently ask whether or not the Wronskian of two polynomials, $p_{n+1}(x)$ of degree $n+1$ and p_n of degree n, determines the zeros of $p_n(x)$ whenever the zeros of the two polynomials are real and interlaced. Clearly, even given the leading coefficients to be 1, the Wronskian does not determine $p_{n+1}(x)$, for the Wronskian is invariant on replacing $p_{n+1}(x)$ by

$$p_{n+1}(x; s) = p_{n+1}(x) - sp_n(x), \tag{1.14}$$

which is equivalent to replacing b_n in (1.10) by $b_n + s$.

There are countless ways in which a positive polynomial may be represented as a sum of squares of polynomials. We might ask what side conditions make the representation unique. The simplest result of this kind is Theorem 1 [1].

THEOREM 1. *A positive polynomial $Q_{2n}(x)$ of exact degree $2n$ has a unique representation*

$$Q_{2n}(x) = \{\pi_n(x)\}^2 + \{\pi_{n-1}(x)\}^2, \tag{1.15}$$

where $\pi_n(x)$ and $\pi_{n-1}(x)$ are real-valued polynomials of degree n and $n-1$, respectively, having (simple) interlaced real zeros.

We only need give the signs of the leading coefficients to make $\pi_n(x)$ and $\pi_{n-1}(x)$ unique. (To accommodate the trivial case, $n=1$, the zeros of $\pi_1(x)$ and $\pi_0(x)$ are said to interlace.) If we suppose the leading coefficients of Q_{2n} and π_n to be 1, and the leading coefficient of π_{n-1} to be positive, then $\pi_n(x)$ and $\pi_{n-1}(x)$ are given simply by

$$\pi_n(x)+i\pi_{n-1}(x)=\prod_{k=1}^{n}(x-z_k), \tag{1.16}$$

where z_k are the zeros of $Q_{2n}(z)$ lying in the lower half-plane. We will prove a similar but more complicated result,

THEOREM 2. *A positive polynomial $Q_{2n}(x)$ of exact degree $2n$ has a unique representation,*

$$Q_{2n}(x)=\sum_{k=0}^{n}\{P_k(x)\}^2 \qquad (-\infty<x<\infty), \tag{1.17}$$

where the P_k are polynomials of exact degree k, real-valued, and orthonormal with respect to a positive measure $d\alpha(x)$ for which Q_{2n} determines the moments through order $2n$.

COROLLARY. *A positive polynomial $Q_{2n}(x)$ of degree $2n\geqq 2$, with leading coefficient* 1, *has the representation*

$$Q_{2n}(x)=p'_{n+1}(x)p_n(x)-p_{n+1}(x)p'_n(x), \tag{1.18}$$

where $p_{n+1}(x)$ and $p_n(x)$ are polynomials of degrees n and $n+1$, respectively, with leading coefficients 1, *and $p_n(x)$ is uniquely determined by the additional condition that the zeros of $p_n(x)$ and $p_{n+1}(x)$ interlace.*

It turns out (Theorem 4) that the additional condition in the corollary can be replaced by the condition that the zeros of either p_{n+1} or p_n be real.

The interpretation of Theorem 2 in terms of the extremal mass function $\rho_n(x)$ is that there is a unique set of n points,

$$-\infty<\lambda_1<\lambda_2<\cdots<\lambda_n<\infty,$$

determined by an (arbitrary) positive polynomial $Q_{2n}(x)$ of (exact) degree $2n$, so that it is possible to find a set of $n+1$ points,

$$-\infty<\gamma_1<\gamma_2<\cdots<\gamma_{n+1}<\infty,$$

such that the two Dirac measures,

$$\sum_{k=1}^{n}\rho_n(\lambda_k)\delta(x-\lambda_k), \qquad \sum_{k=1}^{n+1}\rho_n(\gamma_k)\delta(x-\gamma_k),$$

where

$$\rho_n(x)\equiv\{Q_{2n}(x)\}^{-1},$$

have the same moments through order $2n-1$. (See Fig. 1.) There is a one-parameter family of the latter ("γ") measures, these having the same moments through order $2n$. We would like to see how these point sets may be found, given $\rho_n(x)$ for all real x.

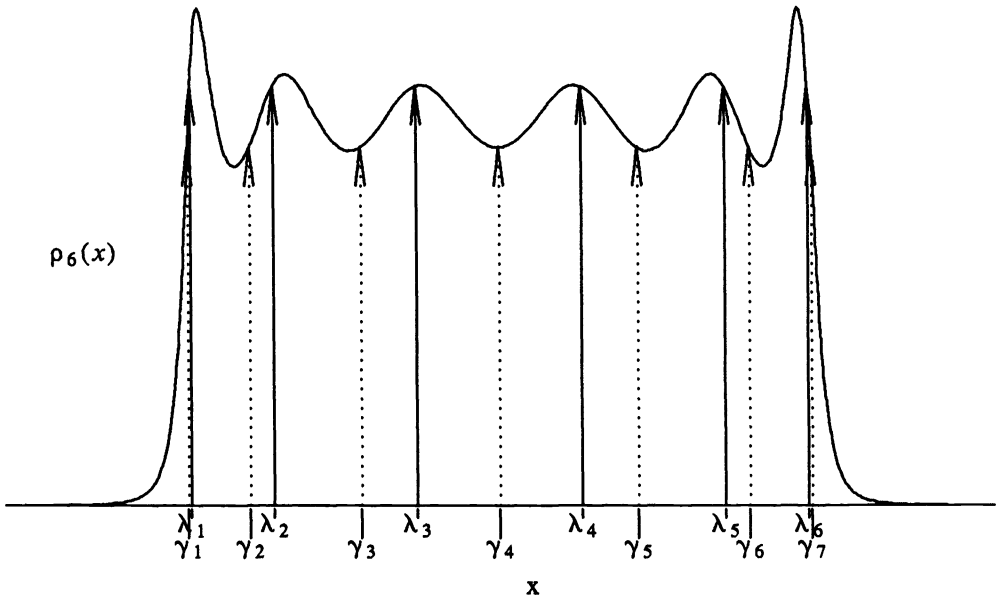


FIG. 1. *Extremal measures determined by a positive polynomial.*

2. The Wronskian of two polynomials. Given $W(x)$, there are countless solutions, f and g, to the equation

$$W(x)=f'(x)g(x)-g'(x)f(x).$$

We can choose, say, $g(x)$, in a quite arbitrary way and then find an f from the first-order differential equation. However, if W is a polynomial and f and g are required to be polynomials, then there are certain constraints on f and g.

We suppose now (only) that $W_n(x)$ is a polynomial of degree $2n$ with leading coefficient 1, and

$$W_n(x)=p'_{n+1}(x)p_n(x)-p'_n(x)p_{n+1}(x), \tag{2.1}$$

where p_{n+1} and p_n are polynomials of degree $n+1$ and n, respectively, with leading coefficients 1. The relation (2.1) may be written

$$W_n(x)=p_n^2(x)\frac{d}{dx}\frac{p_{n+1}(x)}{p_n(x)} \tag{2.2}$$

as well as

$$W_n(x)=-p_{n+1}^2(x)\frac{d}{dx}\frac{p_n(x)}{p_{n+1}(x)}. \tag{2.3}$$

Let us suppose further that p_n and p_{n+1} have simple zeros,

$$p_n(x)=\prod_{k=1}^{n}(x-\lambda_k), \qquad p_{n+1}=\prod_{k=1}^{n+1}(x-\gamma_k). \tag{2.4}$$

Then, remembering that $W_n(x)$ and $p_n^2(x)$ have the same leading coefficient, we see that

$$\frac{W_n(x)}{p_n^2(x)}=1+\sum_1^n\frac{\alpha_k}{(x-\lambda_k)^2}+\sum_1^n\frac{A_k}{(x-\lambda_k)}. \tag{2.5}$$

Since

$$\frac{p_{n+1}(x)}{p_n(x)}=\int^x \frac{W_n(x)}{p_n^2(x)}\,dx = c+x-\sum_1^n \frac{\alpha_k}{x-\lambda_k}+\sum_1^n A_k \log(x-\lambda_k), \tag{2.6}$$

we must have

$$A_k = 0, \qquad k=1,2,\cdots,n. \tag{2.7}$$

Therefore,

$$W_n(x)=p_n^2(x)+\sum_1^n \alpha_k \frac{p_n^2(x)}{(x-\lambda_k)^2}, \tag{2.8}$$

where

$$\alpha_k=\frac{W_n(\lambda_k)}{\{p_n'(\lambda_k)\}^2}.$$

Similarly,

$$W_n(x)=\sum_1^{n+1} \beta_k \frac{p_{n+1}^2(x)}{(x-\gamma_k)^2}, \tag{2.9}$$

where

$$\beta_k=\frac{W_n(\gamma_k)}{\{p_{n+1}'(\gamma_k)\}^2},$$

and the γ_k are the zeros of p_{n+1}. Since the leading coefficient of $W_n(x)$ is 1, we have

$$\sum_1^{n+1} \beta_k = \sum_1^{n+1} \frac{W_n(\gamma_k)}{\{p_{n+1}'(\gamma_k)\}^2}=1. \tag{2.10}$$

Also, from (2.6)-(2.8), we have

$$p_{n+1}(x)=(x+c)p_n(x)-\pi_{n-1}(x), \tag{2.11}$$

where c is an arbitrary constant and

$$\pi_{n-1}(x)=\sum_1^n \frac{W_n(\lambda_k)}{\{p_n'(\lambda_k)\}^2}\frac{p_n(x)}{x-\lambda_k}. \tag{2.12}$$

We have seen that the representation (2.1) of $W_n(x)$ as the Wronskian of two polynomials having leading coefficients 1 and simple zeros implies the representations (2.8) and (2.9) of $W_n(x)$. Now let us see that any one of the representations implies the other two, which will be the case if (2.8) and (2.9) separately imply (2.1).

Suppose first that (2.8) holds. Then

$$\frac{W_n(x)}{p_n^2(x)}=1+\sum_1^n \frac{\alpha_k}{(x-\lambda_k)^2}.$$

Then integrating with respect to x and multiplying by $p_n(x)$ we get

$$p_n(x)\int^x \frac{W_n(x)}{p_n^2(x)}\,dx=(x+c)p_n(x)-\sum_1^n \alpha_k \frac{p_n(x)}{x-\lambda_k}\equiv p_{n+1}(x).$$

Then

$$W_n(x)=p_n^2(x)\frac{d}{dx}\frac{p_{n+1}(x)}{p_n(x)},$$

which is (2.1). Similarly, we get (2.1) from (2.9). Thus we have the following result.

THEOREM 3. *Let $p_{n+1}(x)$ and $p_n(x)$ be polynomials of degree $n+1$ and n, having leading coefficients* 1, *and simple zeros $\{\gamma_k\}$ and $\{\lambda_k\}$, respectively. Then any one of the representations* (2.1), (2.8), *and* (2.9) *implies the other two, and hence we have* (2.11).

This is an interesting addition to Theorems 1 and 2. Now let us make use of additional facts relevant to the original problem.

First we observe that W_n must vanish at common zeros of p_{n+1} and p_n, and as well at multiple zeros of either p_{n+1} or p_n. (Compare (3.5) and (3.6) below.) Now suppose that $W_n(x)$ is positive for all real x, and that the zeros $\{\lambda_k\}$ of p_n are all real. Then the zeros of p_n are real and simple. It follows that the zeros of p_n and the zeros of the polynomial π_{n-1} of degree $n-1$, defined in (2.12), are interlaced, since the sequence

$$\pi_{n-1}(\lambda_k)=\frac{W_n(\lambda_k)}{p_n'(\lambda_k)}, \qquad k=1,2,\cdots,n, \tag{2.13}$$

alternates in sign, on taking $\lambda_{k+1}>\lambda_k$. Furthermore, the leading coefficient of π_{n-1} is positive. Now we have the following lemma [1].

LEMMA. *For $k=0,1,2,\cdots$ let $p_k(x)$ denote a polynomial of degree k in x with leading coefficient* 1. *Now suppose for some integer $n\geqq 1$ that*

$$p_{n+1}(x)=(x-b_n)p_n(x)-a_np_{n-1}(x),$$

where a_n and b_n are real numbers. Then the zeros of p_{n+1} and p_n are real, simple, and interlaced if, and only if, the zeros of p_n and p_{n-1} are real, simple, and interlaced, AND $a_n>0$.

We have seen that p_{n+1} in (2.1) must be related to p_n and π_{n-1} as in (2.11),

$$p_{n+1}(x)=(x+c)p_n(x)-\pi_{n-1}(x),$$

where c is an arbitrary constant of integration. Thus, if in addition to the previous assumptions, p_{n+1} is real-valued on the real axis, then c must be a real number, and therefore, according to the lemma, the zeros of p_{n+1} and p_n are real and interlaced. We arrive at the same conclusion by assuming that the zeros of p_{n+1}, rather than those of p_n, are real.

Assume now in (2.1) that $W_n(x)$ is positive for all real x and that the zeros $\{\gamma_k\}$ of p_{n+1} are all real and therefore simple. Then in (2.9) the β_k are positive, and we have from (2.3)

$$\frac{p_n(x)}{p_{n+1}(x)}=c'+\sum_1^{n+1}\frac{\beta_k}{x-\gamma_k}, \tag{2.14}$$

where the constant of integration c' must vanish in order for p_n to be a polynomial of degree n. (Otherwise, this is equivalent to replacing p_n in (2.1) by $p_n+c'p_{n+1}$, which leaves W_n invariant.) Therefore,

$$p_n(x)=\sum_1^{n+1}\frac{W_n(\gamma_k)}{\{p_{n+1}'(\gamma_k)\}^2}\frac{p_{n+1}(x)}{x-\gamma_k}. \tag{2.15}$$

It follows from (2.14) tht the zeros of p_n and p_{n+1} are interlaced. In summary, we have the following theorem.

THEOREM 4. *Suppose W_n is a given polynomial of degree $2n$ with leading coefficient* 1 *and*

$$W_n(x)=p_{n+1}'(x)p_n(x)-p_n'(x)p_{n+1}(x)>0 \qquad (-\infty<x<\infty),$$

where the polynomials

$$p_{n+1}(x)=\prod_1^{n+1}(x-\gamma_k), \qquad p_n(x)=\prod_1^{n}(x-\lambda_k),$$

are real-valued on the real axis. Then either of the statements,

(i) *All the zeros of p_n are real.*

(ii) *All the zeros of p_{n+1} are real.*

implies

(iii) *The zeros of p_n and p_{n+1} are real, simple, and interlaced.*

Our goal is to show that the hypotheses of Theorem 4 and the first statement imply that p_n is unique.

3. Osculatory interpolation and the differential equation. The representations (2.8) and (2.9) are special cases of simple osculatory interpolation, sometimes called simple Hermite or second-order Lagrange interpolation, where one of the sums drops out. In general, simple osculatory interpolation to a polynomial of degree $2n+1$ at the $n+1$ distinct points γ_k takes the form

$$\pi_{2n+1}(x)=\sum_1^{n+1}\pi_{2n+1}(\gamma_k)\psi_k^2(x)+\sum_1^{n+1}L(\gamma_k)(x-\gamma_k)\psi_k^2(x), \tag{3.1}$$

where

$$\psi_k(x)=\frac{p_{n+1}(x)}{(x-\gamma_k)p'_{n+1}(\gamma_k)},$$

$$L(\gamma_k)=\pi'_{2n+1}(\gamma_k)-\frac{p''_{n+1}(\gamma_k)}{p'_{n+1}(\gamma_k)}\pi_{2n+1}(\gamma_k).$$

Applying this formula to $W_n(x)$ and comparing it with (2.9) we see that the second sum drops out, giving

$$W'_n(\gamma_k)-\frac{p''_{n+1}(\gamma_k)}{p'_{n+1}(\gamma_k)}W_n(\gamma_k)=0,\qquad k=1,2,\cdots,n+1. \tag{3.2}$$

The analogous formula for the n distinct points λ_k, applied to $W_n(x)-p_n^2(x)$ and compared to (2.8), shows that

$$W'_n(\lambda_k)-\frac{p''_n(\lambda_k)}{p'_n(\lambda_k)}W_n(\lambda_k)=0,\qquad k=1,2,\cdots,n. \tag{3.3}$$

Indeed, differentiating (2.1) we get

$$W'_n(x)=p''_{n+1}(x)p_n(x)-p''_n(x)p_{n+1}(x). \tag{3.4}$$

Directly from (2.1) we have

$$W_n(\gamma_k)=p'_{n+1}(\gamma_k)p_n(\gamma_k),\qquad k=1,2,\cdots,n+1, \tag{3.5}$$

$$W_n(\lambda_k)=-p'_n(\lambda_k)p_{n+1}(\lambda_k),\qquad k=1,2,\cdots,n. \tag{3.6}$$

Then (3.2) and (3.3) are obtained easily from (3.4)-(3.6). Also, (3.2) and (3.3) follow directly from (2.9) and (2.8), simply by evaluating W'_n at the interpolation points. Thus we may view the recovery problem as one of finding interpolation points so that either (2.8) or (2.9) holds. If the points are not properly chosen then the derivatives will not match at the points. In the case of (2.9) we know that the γ_k are not unique. However, one of the points may be chosen arbitrarily, except that if it happens to coincide with one of the λ_k then one of the points must recede to infinity, the remaining n points, as we must show, being unique. The difficulty is that the uniqueness depends on the

λ_k being real. Otherwise, there are, in general, many such sets, corresponding to the number of possible differential equations satisfied by p_n. However, it is worth recording the fact that the satisfaction of the interpolation conditions, (3.2) or (3.3), is necessary and sufficient to obtain a solution to (2.1), subject only to the zeros being simple.

THEOREM 5. *Under the hypotheses of Theorem* 3, *the equivalent representations in the conclusion are valid if and only if* (3.2) *or* (3.3) *holds.*

According to (3.3), the function (another Wronskian)

$$W_n'(x)p_n'(x) - p_n''(x)W_n(x) \tag{3.7}$$

vanishes at the zeros of $p_n(x)$. This function is seen to be a polynomial of degree $3n-2$ with leading coefficient $n(n+1)$. Therefore

$$W_n'(x)p_n'(x) - p_n''(x)W_n(x) = n(n+1)Q_{2n-2}(x)p_n(x), \tag{3.8}$$

where $Q_{2n-2}(x)$ is a polynomial of degree $2n-2$ with leading coefficient 1. So given W_n, a polynomial of degree $2n$, how many polynomials Q_{2n-2} are there such that (3.8) has a polynomial solution of degree n? Szegö [2, p. 150] points out (see also Forsyth [3, Vol. 4, pp. 165–169]) that Heine [4, vol. 1, pp. 472–479] studied the following more general problem concerning polynomial solutions of second-order differential equations having polynomial coefficients.

PROBLEM. *Let* $A(x)$ *and* $B(x)$ *be given polynomials of degrees* $m+1$ *and* m, *respectively. To determine a polynomial* $C(x)$ *of degree* $m-1$ *such that the differential equation*

$$A(x)\frac{d^2y}{dx^2} + 2B(x)\frac{dy}{dx} + C(x)y = 0 \tag{3.9}$$

has a solution which is a polynomial of preassigned degree n.

In Szegö's words, "Heine asserts that, in general, there are exactly

$$\nu(m, n) = \binom{n+m-1}{n}$$

determinations of $C(x)$ of this kind" [2]. It is not clear what the precise conditions are for the conclusion to hold. Without further qualification, $\nu(m, n)$ must be regarded as an (achievable) upper bound for the number in question. For example, if $n = 1$, then

$$C(x) = -2\frac{B(x)}{x-b},$$

where b is the zero of $y(x)$, requiring $B(b) = 0$. Thus if $B(x)$ has repeated zeros there will be fewer than m determinations of $C(x)$. So we take the difference,

$$\nu(m, n) - \nu(m, n-1) = \binom{n+m-2}{n},$$

to be an upper bound for the number of determinations of $C(x)$ for which the equation has a solution of exact degree n.

Szegö proves Stieltjes' theorem, which applies to the special case in which $A(x)$ and $B(x)$ are of exact degree $m+1$ and m, respectively, and have real interlacing zeros and leading coefficients of the same sign. In this case there are precisely $\nu(m, n)$ determinations of $C(x)$, each determination and the corresponding y (of degree $\leqq n$) having real zeros. This case was considered of primary importance, as it applied to Lamé's generalized equation (cf. [3, Vol. 4, p. 160]). No connection is made with the problem considered here.

Let us note that if f and g are two solutions of (3.8), then

$$(f''g - g''f)W_n = (f'g - g'f)W_n', \tag{3.10}$$

and hence, as we would expect,

$$f'(x)g(x) - g'(x)f(x) = cW_n(x). \tag{3.11}$$

So if f and g are both polynomials of degree n, then the left-hand side is a polynomial of degree $2n-1$, and hence we must have $c=0$, since W_n is a polynomial of degree $2n$ with leading coefficient 1. So for each determination of Q_{2n-2} there is only one polynomial solution of degree n with leading coefficient 1, say $g(x)=p_n(x)$. But then, according to Theorem 5, (3.11) holds for $f(x)=p_{n+1}(x)$ and $c=1$. Therefore $p_{n+1}(x)$ also satisfies (3.8), which then has the general solution

$$c_1 p_{n+1}(x) + c_2 p_n(x).$$

From Heine's result we may have as many as $\binom{3n-2}{n}$ different determinations of Q_{2n-2} and hence as many determinations of p_n (without stipulating the zeros to be real). Lacking any simple criteria for the solution to have all real zeros, the differential equation appears to have little utility in the general problem. However, repeated (complex) zeros in W_n simplify the problem, as in the following important example.

Example. A simple case in which p_n having real zeros and leading coefficient 1 is uniquely determined is

$$W_n(x) = (1+x^2)^n = p_{n+1}'(x)p_n(x) - p_n'(x)p_{n+1}(x). \tag{3.12}$$

The differential equation for $f = c_1 p_{n+1} + c_2 p_n$ is then

$$2nx(1+x^2)^{n-1}\frac{df}{dx} - (1+x^2)^n\frac{d^2f}{dx^2} = n(n+1)Q_{2n-2}(x)f. \tag{3.13}$$

Since we require p_n (and, say, p_{n+1}) to have only real zeros, we must have

$$Q_{2n-2}(x) = (1+x^2)^{n-1}.$$

Then

$$(1+x^2)\frac{d^2f}{dx^2} - 2nx\frac{df}{dx} + n(n+1)f = 0. \tag{3.14}$$

It is readily verified that

$$f = (x+i)^{n+1} \tag{3.15}$$

satisfies (3.14). The real and imaginary parts of f are then linearly independent solutions of (3.13) for the above determination of Q_{2n-2}. Therefore, if in (3.12) p_n is a polynomial of degree n having only real zeros and leading coefficient 1, then

$$p_n(x) = \frac{1}{n+1}\operatorname{Im}(x+i)^{n+1} = \frac{1}{n+1}\sum_{k\geqq 0}(-1)^k\binom{n+1}{2k+1}x^{n-2k}, \tag{3.16}$$

and we may take as a particular realization of p_{n+1},

$$p_{n+1}(x) = \operatorname{Re}(x+i)^{n+1} = \sum_{k\geqq 0}(-1)^k\binom{n+1}{2k}x^{n+1-2k}. \tag{3.17}$$

The λ_k are determined by

$$\sin\{(n+1)\phi(\lambda_k)\} = 0, \tag{3.18}$$

where

$$\phi(x)=\frac{\pi}{2}-\arctan x \tag{3.19}$$

decreases from π to zero as x increases from $-\infty$ to $+\infty$. The γ_k are determined by

$$\cos\{\theta+(n+1)\phi(\gamma_k)\}=0, \qquad -\frac{\pi}{2}<\theta<\frac{\pi}{2}. \tag{3.20}$$

The polynomials in this example represent the simplest case of orthogonal polynomials found explicitly for a special family of reciprocal-polynomial weight functions [5], the weight function in this case being

$$w_n(x)=\frac{1}{(1+x^2)^{n+1}}, \tag{3.21}$$

where the polynomials $p_k(x)$ are proportional to the Jacobi polynomials $P_k^{(-n,-n)}(ix)$. The imaginary part of

$$\int_{-\infty}^{\infty}\left\{\frac{x^n}{(x-i)^{n+1}}+\frac{1}{x-i}\right\}dx=0$$

gives

$$h_n=\int_{-\infty}^{\infty}\frac{x^n p_n(x)\,dx}{(1+x^2)^{n+1}}=\frac{\pi}{n+1}. \tag{3.22}$$

Thus in this case we have a simple relation between the maximal mass function and the weight function,

$$\rho_n(x)=\frac{h_n}{W_n(x)}=\frac{\pi}{n+1}(1+x^2)w_n(x). \tag{3.23}$$

In the general case, it is not clear how Q_{2n-2} may be chosen so that p_n will have only real zeros. We note that

$$n(n+1)Q_{2n-2}(x)=p''_{n+1}(x)p'_n(x)-p''_n(x)p'_{n+1}(x), \tag{3.24}$$

which follows from substituting (2.1) in (3.8). The Wronskian in (3.24) will be positive if the zeros of p'_n and p'_{n+1} interlace, which will be the case if the zeros of p_n and p_{n+1} interlace [1]; i.e., if the zeros of p_n are real and p_{n+1} is real. Hence

$$Q_{2n-2}(x)>0, \qquad -\infty<x<\infty, \tag{3.25}$$

is a necessary condition for p_n to have only real zeros. However, the condition is not sufficient, as shown by the example

$$p_4(x)=(x^2-\tfrac{1}{3})(x^2+1),$$
$$p_5(x)=(x^3-3x)(x^2+1),$$
$$W_4(x)=p'_5(x)p_4(x)-p'_4(x)p_5(x)=(x^2+1)^4,$$
$$20Q_6(x)=p''_5(x)p'_4(x)-p''_4(x)p'_5(x)=(x^2+1)^2(20x^2+4).$$

Although it is not clear from the differential equation, we are able to see from physical considerations that there is at least one polynomial p_n of degree n having only real zeros, and leading coefficient 1, such that (2.1) holds for an arbitrary positive polynomial W_n of degree $2n$ with leading coefficient 1. The uniqueness of p_n is then established by an induction argument applied to a pair of minimization problems.

4. The electrostatic equilibrium problems. The interpolation conditions (3.2) and (3.3) which lead to the differential equation define two related electrostatic equilibrium problems, which we designate by $[W_n; n]$ and $[W_n; n+1]$.

Given a positive polynomial $W_n(x)$ of degree $2n$ with leading coefficient 1, solve the following problems:

PROBLEM $[W_n; n]$. *Find a set* $\{\lambda_k\}$,

$$-\infty < \lambda_1 < \lambda_2 < \cdots < \lambda_n < \infty,$$

such that

$$\frac{W_n'(\lambda_k)}{W_n(\lambda_k)} = \frac{p_n''(\lambda_k)}{p_n'(\lambda_k)}, \qquad k = 1, 2, \cdots, n, \tag{4.1}$$

where

$$p_n(x) = \prod_{k=1}^{n} (x - \lambda_k).$$

PROBLEM $[W_n; n+1]$. *Find a set* $\{\gamma_k\}$,

$$-\infty < \gamma_1 < \gamma_2 < \cdots < \gamma_{n+1} < \infty,$$

such that

$$\frac{W_n'(\gamma_k)}{W_{n+1}(\gamma_k)} = \frac{p_{n+1}''(\gamma_k)}{p_{n+1}'(\gamma_k)}, \qquad k = 1, 2, \cdots, n+1, \tag{4.2}$$

where

$$p_{n+1}(x) = \prod_{k=1}^{n+1} (x - \gamma_k).$$

Now write

$$p_n(x) = \prod_{k=1}^{n} (x - \lambda_k) = (x - \lambda_m)\pi_{n-1}(x; m),$$

$$p_n'(x) = (x - \lambda_m)\pi_{n-1}'(x) + \pi_{n-1}(x),$$

$$p_n''(x) = (x - \lambda_m)\pi_{n-1}''(x) + 2\pi_{n-1}'(x).$$

Then

$$\frac{p_n''(\lambda_m)}{p_n'(\lambda_m)} = 2\frac{\pi_{n-1}'(\lambda_m)}{\pi_{n-1}(\lambda_m)} = 2\sum_k{}' \frac{1}{\lambda_m - \lambda_k}. \tag{4.3}$$

Thus (4.1) may be written

$$\frac{1}{2}\frac{W_n'(\lambda_m)}{W_n(\lambda_m)} = \sum_k{}' \frac{1}{\lambda_m - \lambda_k}, \qquad m = 1, 2, \cdots, n \quad (-\infty < \lambda_1 < \lambda_2 < \cdots < \lambda_n < \infty). \tag{4.4}$$

This says that the repelling force on a unit negative charge at the point λ_m due to $n-1$ unit negative charges at the other λ_k should be balanced by the attracting force due to $2n$ positive charges of $\{\frac{1}{2}\}$ unit each held at the $2n$ zeros of W_n, $x_k \pm iy_k$, $k = 1, 2, \cdots, n$, when the negative charges are free to move on the real axis. There will be at least one such equilibrium point, since the repelling force assures the separation of the negative charges and the excess of one unit positive charge assures that the extreme right and left negative charges will not recede to infinity. So one equilibrium point could be

found by positioning the negative charges at n distinct points on the real axis and then letting them move with a velocity proportional to the force at their respective positions. Since the zeros of W_n occur in conjugate pairs, this force will always be directed along the real axis.

The problem $[W_n; n+1]$ has a similar interpretation, except that now there are $n+1$ negative charges free to move on the real axis, and we know in this case that there is not a unique equilibrium point. However, we now know from the interpretation of the problem $[W_n; n]$, and the previous results, that the problem $[W_n; n+1]$ has at least a one-parameter family of solutions.

THEOREM 6. *To each positive polynomial $W_n(x)$ of degree $2n$ with leading coefficient 1 there corresponds at least one polynomial $p_n(x)$ of degree n with leading coefficient 1, having only real simple zeros $\{\lambda_k\}$, such that*

$$W_n(x) = p'_{n+1}(x)p_n(x) - p'_n(x)p_{n+1}(x),$$

where p_{n+1} is a polynomial of degree $n+1$ with leading coefficient 1, partially determined by p_n according to

$$p_{n+1}(x) = (x+c)p_n(x) - a_n p_{n-1}(x),$$

where c is an arbitrary constant, and p_{n-1} is a polynomial of degree $n-1$ with leading coefficient 1 determined by p_n according to

$$a_n p_{n-1}(x) = \sum_{k=1}^{n} \frac{W_n(\lambda_k)}{\{p'_n(\lambda_k)\}^2} \frac{p_n(x)}{x-\lambda_k},$$

thereby determining real numbers b_k, positive numbers a_k (except a_0 which is arbitrary), and polynomials $p_k(x)$ of degree k with leading coefficients 1 such that

$$xp_k(x) - p_{k+1}(x) - b_k p_k(x) = a_k p_{k-1}(x), \qquad k = 0, 1, \cdots, n-1,$$

and hence each such polynomial $p_n(x)$ completely determines the representation

$$W_n(x) = p_n^2(x) + a_n W_{n-1}(x),$$

where

$$W_k(x) = p_k^2(x) + a_k W_{k-1}(x), \qquad k = 1, 2, \cdots, n.$$

Next we show that there is only one such polynomial $p_n(x)$.

5. The paired extremal problems. To prove that the equilibrium (interpolation) problem $[W_n; n]$ has a unique solution it seems necessary to consider a pair of extremal problems, $[W_n; n]_0$ and $[W_n; n+1]_0$.

Given $W_n(x)$, a positive polynomial of degree $2n$ with leading coefficient 1, solve the following problems:

PROBLEM $[W_n; n]_0$. *Find a set $\{\lambda_k\}$,*

$$-\infty < \lambda_1 < \lambda_2 < \cdots < \lambda_n < \infty,$$

that gives

$$E_0(W_n; n) \equiv \min_{p_n} E(W_n; p_n), \tag{5.1}$$

where

$$E(W_n; p_n) = \frac{1}{2} \sum_{k=1}^{n} \log \left\{ \frac{W_n(\lambda_k)}{p'_n(\lambda_k)} \right\}^2 \tag{5.2}$$

and

$$p_n(x)=\prod_{k=1}^{n}(x-\lambda_k).$$

PROBLEM $[W_n; n+1]_0$. *Find a set* $\{\gamma_k\}$,

$$-\infty<\gamma_1<\gamma_2<\cdots<\gamma_{n+1}<\infty,$$

that gives

$$E_o(W_n; n+1)\equiv\min_{p_{n+1}} E(W_n; p_{n+1}), \tag{5.3}$$

where

$$E(W_n; p_{n+1})=\frac{1}{2}\sum_{k=1}^{n+1}\log\left\{\frac{W_n(\gamma_k)}{p'_{n+1}(\gamma_k)}\right\}^2 \tag{5.4}$$

and

$$p_{n+1}(x)=\prod_{k=1}^{n+1}(x-\gamma_k).$$

Remark. It is simply a matter of convenience to take the leading coefficient of W_n to be 1. We have for a positive number A,

$$E(AW_n; pn)=n\log A+E(W_n; p_n),$$

$$E(AW_n; p_{n+1})=(n+1)\log A+E(W_n; p_{n+1}).$$

Now suppose we vary λ_m in (5.2). From (4.3) we find

$$\begin{aligned}&\frac{\partial}{\partial\lambda_m}p'_n(\lambda_m)=\frac{1}{2}p''_n(\lambda_m)\\ &\frac{\partial}{\partial\lambda_m}p'_n(\lambda_k)=-\frac{p'_n(\lambda_k)}{\lambda_k-\lambda_m},\qquad k\neq m.\end{aligned} \tag{5.5}$$

Then setting

$$0=\frac{\partial}{\partial\lambda_m}E(W_n; p_n)=\frac{W'_n(\lambda_m)}{W_n(\lambda_m)}-\frac{1}{2}\frac{p''_n(\lambda_m)}{p'_n(\lambda_m)}+\sum_k{}'\frac{1}{\lambda_k-\lambda_m}$$

for $m=1, 2, \cdots, n$, we get, using (4.3), problem $[W_n; n]$. Similarly, setting

$$\frac{\partial}{\partial\gamma_m}E(W_n; p_{n+1})=0,\qquad m=0, 1, 2, \cdots, n+1,$$

we get problem $[W_n; n+1]$. So solutions of the minimization problems are by necessity solutions of the corresponding electrostatic equilibrium problems.

Now suppose that p_n is some solution of the equilibrium problem $[W_n; n]$. Then (cf. Theorem 6) paired with p_n is the corresponding one-parameter family of solutions of the problem $[W_n; n+1]$,

$$p_{n+1}(x; b)=(x-b)p_n(x)-a_np_{n-1}(x),\qquad -\infty<b<\infty. \tag{5.6}$$

Since b is real, the zeros $\gamma_k=\gamma_k(b)$ of p_{n+1} are real and simple. Let us see that $E(W_n; p_{n+1})$ is independent of the parameter b for the equilibrium solutions (5.6).

We have from the Wronskian representation (2.1)

$$p'_{n+1}(\gamma_k(b); b) = \frac{W_n(\gamma_k(b))}{p_n(\gamma_k(b))}. \tag{5.7}$$

Then

$$E(W_n; p_{n+1}) = \frac{1}{2}\sum_{k=1}^{n+1} \log\{p_n^2(\gamma_k(b))\} \tag{5.8}$$

and

$$\frac{d}{db}E(W_n; p_{n+1}) = \sum_{k=1}^{n+1} \gamma'_k(b)\frac{p'_n(\gamma_k(b))}{p_n(\gamma_k(b))}. \tag{5.9}$$

Writing

$$p_{n+1}(x; b) = p_{n+1}(x; 0) - bp_n(x), \tag{5.10}$$

we have

$$\frac{d}{db}\{p_{n+1}(\gamma_k(b); b)\} = \{p'_{n+1}(\gamma_k(b); 0) - bp'_n(\gamma_k(b))\}\gamma'_k(b) - p_n(\gamma_k(b)) = 0$$

or

$$\gamma'_k(b) = \frac{p_n(\gamma_k(b))}{p'_{n+1}(\gamma_k(b); b)}. \tag{5.11}$$

Thus we find

$$\frac{d}{db}E(W_n; p_{n+1}) = \sum_{k=1}^{n+1} \frac{p'_n(\gamma_k(b))}{p'_{n+1}(\gamma_k(b); b)} = 0, \tag{5.12}$$

since

$$0 = \lim_{r\to\infty} \int_{|z|=r} \frac{\pi_{n-1}(z)}{p_{n+1}(z)}\,dz = 2\pi i \sum_{k=1}^{n+1} \frac{\pi_{n-1}(\gamma_k)}{p'_{n+1}(\gamma_k)}$$

for any polynomial π_{n-1} of degree $n-1$. Therefore we may evaluate the constant $E(W_n; p_{n+1})$ by letting $b \to \infty$:

$$E(W_n; p_{n+1}) = \lim_{b\to\infty} \sum_{k=1}^{n+1} \log|p_n(\gamma_k(b))|. \tag{5.13}$$

We have

$$0 = p_{n+1}(\gamma_k(b); b) = (\gamma_k(b) - b)p_n(\gamma_k(b)) - a_n p_{n-1}(\gamma_k(b))$$

or

$$p_n(\gamma_k(b)) = \frac{a_n p_{n-1}(\gamma_k(b))}{\gamma_k(b) - b}.$$

Now as $b \to \infty$ the first n zeros of p_{n+1} will tend to the zeros $\{\lambda_k\}$ of p_n, while the largest zero will differ slightly from b. We have

$$p_n(\gamma_k(b)) = \frac{-a_n p_{n-1}(\lambda_k)}{b} + O(b^{-2}), \qquad k = 1, 2, \cdots, n,$$

$$p_n(\gamma_{n+1}(b)) = b^n + O(b^{n-1}).$$

The last term in (5.13) then cancels the $\log|b|$ appearing in the first n terms. Thus

$$E(W_n; p_{n+1}) = \sum_{k=1}^{n} \log|a_n p_{n-1}(\lambda_k)|. \tag{5.14}$$

But we have from (5.6)

$$a_n p_{n-1}(\lambda_k) = -p_{n+1}(\lambda_k)$$

and from (2.1)

$$-p_{n+1}(\lambda_k)p'_n(\lambda_k) = W_n(\lambda_k)$$

or (cf. (2.13))

$$a_n p_{n-1}(\lambda_k) = \frac{W_n(\lambda_k)}{p'_n(\lambda_k)}.$$

Thus for a pair of equilibrium solutions, p_n and p_{n+1}, we have

$$E(W_n; p_{n+1}) = \sum_{k=1}^{n} \log\left|\frac{W_n(\lambda_k)}{p'_n(\lambda_k)}\right| = E(W_n; p_n). \tag{5.15}$$

Uniqueness. If p_n is a solution of the problem $[W_n; n]$, then

$$W_n(x) = p_n^2(x) + a_n W_{n-1}(x), \tag{5.16}$$

where

$$W_{n-1}(x) = p'_n(x)p_{n-1}(x) - p'_{n-1}(x)p_n(x) \tag{5.17}$$

and

$$a_n p_{n-1}(x) = \sum_{k=1}^{n} \frac{W_n(\lambda_k)}{\{p'_n(\lambda_k)\}^2} \frac{p_n(x)}{x-\lambda_k}. \tag{5.18}$$

We wish to show that p_n is the unique solution of the problem $[W_n; n]_0$ by virtue of being a solution of the problems $[W_n; n]$ and $[W_{n-1}; n]$.

Consider a competitor π_n in the problem $[W_n; n]_0$,

$$\pi_n(x) = \prod_{k=1}^{n} (x - t_k), \qquad -\infty < t_1 < t_2 < \cdots < t_n < \infty.$$

Since

$$W_n(t_k) = p_n^2(t_k) + a_n W_{n-1}(t_k) \geqq a_n W_{n-1}(t_k), \tag{5.19}$$

we have

$$\begin{aligned} E(W_n; \pi_n) &= \frac{1}{2}\sum_{k=1}^{n} \log\left\{\frac{W_n(t_k)}{\pi'_n(t_k)}\right\}^2 \\ &\geqq n \log a_n + \frac{1}{2}\sum_{k=1}^{n} \log\left\{\frac{W_{n-1}(t_k)}{\pi'_n(t_k)}\right\}^2, \end{aligned} \tag{5.20}$$

and hence

$$E(W_n; \pi_n) \geqq n \log a_n + E(W_{n-1}; \pi_n). \tag{5.21}$$

Thus π_n is confronted with the problem $[W_{n-1}; n]_0$, where W_{n-1} is determined by its competitor p_n, a particular solution of the problem $[W_n; n]$.

Now suppose the following.

Induction Hypothesis (I.H.). The problem $[W_m; m]$ has a unique solution for $m = 0, 1, \cdots, n-1$.

The problem $[W_0; 0]$ trivially has a unique solution. Also the problem $[W_1; 1]$ described by

$$W_1(x) = x^2 + A_1 x + A_2 = p_1^2(x) + a_1 > 0,$$

where

$$p_1(x) = x - \lambda_1, \qquad -\infty < \lambda_1 < \infty,$$

clearly has a unique solution, $\lambda_1 = -A_1/2$, which is, in fact, the simplest case of Theorem 1 stated in the Introduction. So I.H. is true for $n = 2$.

Now since p_{n-1} defined in (5.18) is, according to (5.17), a solution of the problem $[W_{n-1}; n-1]$, it is, as hypothesized, the unique solution of that problem and hence the unique solution of the problem $[W_{n-1}; n-1]_0$. It follows from Theorem 6 that the problem $[W_{n-1}; n]$ has the "semi-unique" solution

$$p_n(x) = (x - b_{n-1})p_{n-1}(x) - a_{n-1}p_{n-2}(x), \qquad -\infty < b_{n-1} < \infty. \tag{5.22}$$

Therefore the problem $[W_{n-1}; n]_0$ also has the semi-unique solution (5.22). Hence,

$$E_o(W_{n-1}; n) = E(W_{n-1}; p_n) = E(W_{n-1}; p_{n-1}) = E_o(W_{n-1}; n-1). \tag{5.23}$$

Therefore in (5.21) we have

$$E(W_{n-1}; \pi_n) \geqq E(W_{n-1}; p_n) = E_o(W_{n-1}; n-1), \tag{5.24}$$

and hence

$$E(W_n; \pi_n) \geqq n \log a_n + E(W_{n-1}; p_n). \tag{5.25}$$

Now in order for equality to hold in (5.25) we must first have equality in (5.24), which requires π_n to be a semi-unique solution of the problem $[W_{n-1}; n]$; i.e.,

$$\pi_n(x) = p_n(x) - cp_{n-1}(x).$$

Next, equality must hold in (5.21), requiring in (5.19)

$$p_n(t_k) = 0, \qquad k = 1, 2, \cdots, n, \qquad \pi_n(x) \equiv p_n(x).$$

Note that b_{n-1} in the semi-unique solution (5.22) of the problem $[W_{n-1}; n]$ is determined by p_n being a solution of the problem $[W_n; n]$ so as to give the required coefficient of x^{2n-1} in

$$W_n(x) = x^{2n} + A_1 x^{2n-1} + \cdots = p_n^2(x) + a_n(x^{2n-2} + \cdots).$$

Hence the problem $[W_n; n]$ has a unique solution provided the problem $[W_{n-1}; n-1]$ has a unique solution. Therefore the induction hypothesis is true for all $n \geqq 1$, and we have proved Theorem 2 as stated in the Introduction, where the squares of the orthonormal polynomials

$$P_k^2(x) = \frac{p_k^2(x)}{h_k}$$

are unique.

Also we have

$$E_o(W_n; n+1) = E_o(W_n; n) = n \log a_n + E_o(W_{n-1}; n-1), \tag{5.26}$$

and therefore

$$E_o(W_n; n+1) = E_o(W_n; n) = \sum_{k=1}^{n} k \log a_k, \tag{5.27}$$

where the a_k (the recursion coefficients in (1.7)) are uniquely determined by a (semi-unique) solution of the problem $[W_n; n+1]$ or by the unique solution of the problem $[W_n; n]$. Thus $E_o(W_n; n)$ is the logarithm of the product of the coefficients of p_k^2 in the sum

$$W_n(x) = p_n^2(x) + a_n p_{n-1}^2(x) + a_n a_{n-1} p_{n-2}^2(x) + \cdots + a_n a_{n-1} \cdots a_1. \tag{5.28}$$

6. Bounds for $E_o(W_n; n)$. Bounds for the minimum defined in (5.1) may be obtained from the minimum for the special case

$$W_n(x) = (x^2+1)^n. \tag{6.1}$$

In general, if p_n is the solution of the problem $[W_n; n]$ then

$$E_o(W_n; n) = \sum_{k=1}^{n} \log \left| \frac{W_n(\lambda_k)}{p_n'(\lambda_k)} \right| \tag{6.2}$$

and

$$W_n(x) = p_{n+1}'(x) p_n(x) - p_n'(x) p_{n+1}(x).$$

Thus (6.2) may be written

$$E_o(W_n; n) = \sum_{k=1}^{n} \log |p_{n+1}(\lambda_k)|. \tag{6.3}$$

From the example in § 4 we have for the special case (6.1)

$$p_n(x) = \frac{1}{2i(n+1)} \{(x+i)^{n-1} - (x-i)^{n+1}\} \tag{6.4}$$

and

$$p_{n+1}(x) + i(n+1) p_n(x) = (x+i)^{n+1}. \tag{6.5}$$

Then

$$p_{n+1}(\lambda_k) = (\lambda_k + i)^{n+1}. \tag{6.6}$$

From (6.4) we have

$$(\lambda_k + i)^{n+1} = (\lambda_k - i)^{n+1}$$

or (ignoring the ordering requirement)

$$\frac{\lambda_k + i}{\lambda_k - i} = e^{i2\pi k/(n+1)},$$

which gives

$$\lambda_k = \frac{1}{\tan k\pi/(n+1)}, \qquad k = 1, 2, \cdots, n \tag{6.7}$$

$$\lambda_k + i = \frac{e^{ik\pi/(n+1)}}{\sin k\pi/(n+1)} \tag{6.8}$$

$$p_{n+1}(\lambda_k) = \frac{(-1)^k}{\{\sin k\pi/(n+1)\}^{n+1}}. \tag{6.9}$$

Defining

$$\sigma_n = E_o((x^2+1)^n;\, n) \tag{6.10}$$

we have from (6.3) and (6.9)

$$\sigma_n = -(n+1)\sum_{k=1}^{n} \log\left\{\sin\frac{k\pi}{n+1}\right\}. \tag{6.11}$$

From the factorization

$$\sin(n+1)t = 2^n \sin t \prod_{k=1}^{n} \sin(t + k\pi/(n+1))$$

we have

$$\prod_{k=1}^{n} \sin\frac{k\pi}{n+1} = \frac{n+1}{2^n}. \tag{6.12}$$

Hence

$$\sigma_n = (n+1)\log\frac{2^n}{n+1}. \tag{6.13}$$

In connection with the general expression (5.27) for $E_o(W_n;\, n)$, the recursion coefficients for the special case (6.1) are found to be [5]

$$a_k = a_k(n) = \frac{k(2n+2-k)}{(2n+1-2k)(2n+3-2k)}, \qquad k = 1, 2, \cdots, n. \tag{6.14}$$

Thus we have the rather interesting sum

$$\sum_{k=1}^{n} k \log\frac{k(2n+2-k)}{(2n+1-2k)(2n+3-2k)} = (n+1)\log\frac{2n}{n+1}. \tag{6.15}$$

By appropriate scaling of the solution of the special problem we find that

$$E_o((x^2+a^2)^n;\, n) = \frac{n(n+1)}{2}\log a^2 + \sigma_n, \qquad (a^2 > 0). \tag{6.16}$$

In the general case we have

$$W_n(x) = \prod_{j=1}^{n} \{(x-x_j)^2 + y_j^2\}. \tag{6.17}$$

Then writing

$$\sum_{k=1}^{n} \log\frac{W_n(\lambda_k)}{|p_n'(\lambda_k)|} = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \log\frac{\{(x-x_j)^2+y_j^2\}^n}{|p_n'(\lambda_k)|}$$

we see from (6.16) that

$$E_o(W_n;\, n) \geqq \frac{n+1}{2}\sum_{j=1}^{n} \log y_j^2 + \sigma_n. \tag{6.18}$$

In the other direction, we have, since the average of the logarithms cannot exceed the logarithm of the average,

$$\frac{1}{n}\log W_n(x) \leqq \log\{(x-\bar{x})^2 + c^2\},$$

and hence

$$E_o(W_n;\, n) \leqq \frac{n(n+1)}{2}\log c^2 + \sigma_n, \tag{6.19}$$

where

$$c^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j^2+y_j^2) - \left\{\frac{1}{n}\sum_{j=1}^{n} x_j\right\}^2 .$$

REFERENCES

[1] B. F. LOGAN, *The basic Hurwitz test*, Tech. Mem., Nov. 14, 1986.
[2] GABOR SZEGÖ, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Pub. 23, American Mathematical Society, Providence, RI, 1959.
[3] ANDREW RUSSELL FORSYTH, *Theory of Differential Equations*, Six volumes bound as three, Dover, New York, 1959.
[4] E. HEINE, *Handbuch der Kugelfunctionen*, Vols. I and II, Second edition, Berlin, 1878, 1881.
[5] B. F. LOGAN, *The Symmetric Jacobi-Theta Polynomials*, Tech. Mem., Nov. 14, 1986.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 1051–1064, July 1990
©1990 Society for Industrial and Applied Mathematics
013

STABLE POSITIVITY OF POLYNOMIALS OBTAINED FROM THREE-TERM DIFFERENCE EQUATIONS*

D. E. HANDELMAN†

Abstract. Let $\{u_0 = 1, u_1 = x, u_2, u_3, \cdots\}$ be a sequence of polynomials in one variable x defined recursively via

$$u_{n+1} = (u_1 - a_n)u_n - b_n u_{n-1} \qquad \text{for } n > 1,$$

where $\{a_1, a_2, \cdots\}$ and $\{b_1, b_2, \cdots\}$ are sequences of real numbers and suppose that any product $u_k \cdot u_n$ is expressible as a positive (real) combination of the u_i's. Let f be a real polynomial in x, and let u be a positive combination of the u_i's. consider the problem of deciding whether there exists an integer n such that the product of polynomials, $u^n f$, is a positive combination of the u_i's. If the sequences of a's and b's are not too degenerate, the solution depends only on f being positive as a function on a real interval of the form $[M, \infty)$ for some real M determined by the sequences.

Key words. orthogonal polynomials, three-term recurrence relation, dimension group, pure state, Dedekind domain

AMS(MOS) subject classifications. 42C05, 13J25

0. Introduction.

Let $\{u_0 = 1, u_1 = x, u_2, u_3, \cdots\}$ be a sequence of polynomials in one variable x defined recursively via

$$u_{n+1} = (u_1 - a_n)u_n - b_n u_{n-1} \qquad \text{for } n > 1,$$

where $\{a_1, a_2, \cdots\}$ and $\{b_1, b_2, \cdots\}$ are sequences of real numbers; a normalization has been imposed, wherein $a_0 = 1$ and $b_0 = 0$. Assume that any product $u_k u_n$ is expressible as a positive (real) combination of the u_i's. A sufficient condition for this to occur is given in [AI, 2.9]; it is simply that

$$0 \le a_1 \le a_2 \le \cdots \quad \text{and} \quad 0 \le b_1 \le b_2 \le \cdots.$$

Let f be a real polynomial in x, and let u be a positive combination of the u_i's. We consider the following problem:

(*) Decide if there exists an integer n such that the product of polynomials, $u^n f$, is a positive combination of the u_i's.

Let M be the real number or ∞ defined by means of

$$M = \sup a_i + 2\sqrt{\sup b_i}.$$

The positivity condition on the products $u_k u_n$ forces all of the a's and b's to be nonnegative, so that M is well-defined as an element of $\mathbf{R} \cup \{\infty\}$. Suppose that the sequences $\{a_i\}$, $\{b_i\}$ are monotone nondecreasing and for all $n \ge 1$, all of the b_n are nonzero, and either of the following hold:

(A) $a_1 \ne 0$ *or*

(B) $u = \sum_{i=0}^{k} \alpha_i u_i$ where the real numbers α_i are strictly positive real numbers.

Then such an integer n exists if and only if

*Received by the editors August 26, 1985; accepted for publication (in revised form) July 25, 1989. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

† Mathematics Department, University of Ottawa, Ottawa, Ontario, Canada K1N 6N5, email, dehsg@uottawa.bitnet.

(i) The leading coefficient of f is positive when $M = \infty$.
(ii) Viewed as a function, f is strictly positive on the interval $[M, \infty)$ when $M < \infty$.

This is the content of Theorems 4.1 and 4.5. Even allowing some b's to be zero (but assuming monotonicity of the sequences $\{a_i\}$, $\{b_i\}$), necessary and sufficient conditions on f are determined in Theorem 5.1; however, these are considerably more complicated. In fact, if condition (B) above holds, the monotonicity hypothesis on the sequences can be weakened to some of its consequences.

There are a few special cases of this type of result in the literature. For example, if $a_i = b_j = 0$ for all i and j greater than zero, then $u_i = x^i$, and results in this direction were obtained as early as 1883 by Poincaré [Po]. This situation is not covered by either conditions (A) or (B), but a complete solution is given in [H2]. The case in which $a_i = b_j = 1$ for all i and j arose in connection with the classification of actions of a special type of the group SO(3) on a class of C*-algebras, and is equivalent to a random walk problem on $\mathbf{Z}$ [H1; §§I and II]. If $a_i = 0$ and $b_j = 1$, the same connections occur, this time with SU(2) instead of SO(3), and it is possible to reduce to SO(3). The techniques employed here evolved from those developed to deal with the questions considered in [H1]. My original approach, using estimates of the asymptotic behaviour of coefficients in products of large numbers of polynomials, never seemed to yield very good results, hence the somewhat algebraic and functional analytic flavour of the techniques used here.

1. Preliminaries. Fix the sequences of nonnegative real numbers, $\{a_i\}$ and $\{b_i\}$, and define the corresponding sequence of polynomials, $\{u_i\}$; these are orthogonal with respect to a suitable measure on $\mathbf{R}$ (if no b_i is zero). This set of polynomials is a basis for the polynomial algebra, $\mathbf{R}[x]$, in the variable x. If we assume that for all k and n, the products $u_k u_n$ are nonnegative linear combinations of the u_i's, then $\mathbf{R}[x]$ becomes a partially ordered ring, with positive cone,

$$\mathbf{R}[x]^+ = \left\{ \sum \lambda_i u_i \mid \lambda_i = 0 \text{ a.e., and } \lambda_i \in \mathbf{R}^+ \right\} .$$

Let u be an element of $\mathbf{R}[x]^+$. We wish to describe the set

$$\{ f \in \mathbf{R}[x] \mid \text{there exists } m \text{ such that } u^m f \text{ belongs to } \mathbf{R}[x]^+ \}$$

or at least decide whether a specific polynomial belongs to it. We follow the construction in [H1, §I], to set up the framework for this to be solved. Define the partially ordered ring $S_u = \mathbf{R}[x, u^{-1}]$ consisting of rational functions whose denominators are powers of u, with positive cone,

$$S_u^+ = \{ f u^{-k} \mid f \in \mathbf{R}[x];\ k \in \mathbf{N}, \text{ and there exists } m \text{ such that } u^m f \in \mathbf{R}[x]^+ \} .$$

Next, we define the order ideal of S_u generated by 1; this is not an ideal in the ring sense, but in the sense of partially ordered abelian groups (the smallest convex subgroup of S_u containing 1, and generated by the positive elements that it contains). Set

$$R_u = \{ s \in S_u \mid \text{there exists an integer } N \text{ such that } -N \leq s \leq N \text{ holds in } S_u \} .$$

Equipped with the ordering it inherits from S_u (the "relative ordering"), it is routine to verify that R_u is a partially ordered subring of S_u, and that 1 is an *order unit* of R_u, namely that for all r in R_u, there exists an integer N with $r \leq N$ in R_u.

If G is a partially ordered abelian group (for example, the additive group underlying R_u), then we say that G is *unperforated* if whenever m is a positive integer and g is an element of G such that $mg \geq 0$, then $g \geq 0$. This clearly holds for each of $\mathbf{R}[x]$,

R_u, and S_u. Let (G, h) be a partially ordered abelian group with order unit h; a *state* is an order-preserving group homomorphism $\alpha:(G, h) \to (\mathbf{R}, 1)$. If no order unit exists, or if one exists but is not referred to, we will also call any nonzero order-preserving group homomorphism from G to $\mathbf{R}$ a state as well, possibly prefixed *unnormalized.* A state is *pure* (or extremal) if it cannot be expressed as a nontrivial convex linear combination of other states. Our main tool in proving the advertised results is the following, a combination of results in [H1].

THEOREM 1.1. [H1; I.1 & I.2] *Let G be a partially ordered ring admitting an order unit.*

(a) *Suppose G is unperforated, and let g be an element of G. Suppose that for all multiplicative states α, $\alpha(g) > 0$ (strictly). Then g is an order unit, and thus is positive in G.*

(b) *If G is a partially ordered ring with 1 as an order unit, then every pure state of $(G, 1)$ is a multiplicative homomorphism from G to $\mathbf{R}$.*

Theorem 1.1 applies to R_u, and our first step will be to determine the pure states on R_u. It turns out that they are either evaluations at certain points of $\mathbf{R}$ (regarding R_u as a set of rational functions of x), or their limit at infinity. After the pure states are determined, some crucial algebraic properties of R_u are established. For example, if condition (B) of the Introduction holds, then R_u is a Dedekind domain. The unique factorization of ideals that thereby results is used to find all of the order ideals that are ideals. Then Theorem 1.1 can be applied to these order ideals. The determination of the positive cone of R_u then follows by applying Theorem 1.1(a) to the smallest order ideal containing the element. When (A) holds, the problem can be reduced to the situation in which (B) holds.

2. States of R_u. In this section, we determine the pure states of R_u; under appropriate hypotheses on u, the pure states consist of point evaluations $r \mapsto r(t)$ for any real number $t \geq M$ (M as defined in the Introduction), together with an additional discrete set of positive values of t—no such need exist—and the state $r \mapsto \lim_{t\to\infty} r(t)$ which always exists, by l'Hôpital's rule. Equipped with the point-open topology (as a set of functions on $\mathbf{R}$), the pure state space is thus the union of a possibly empty discrete set with either the one-point compactification of a half-line, or just the singleton $\{\infty\}$. We require a number of straightforward lemmas about sign changes among the u_i. Our basic reference is the first fifteen pages of [AI].

LEMMA 2.1. *Suppose there exists j such that for all $k \geq j$, $a_k = K > 0$ and $b_k = L > 0$. If β is a real number strictly less than $K + 2\sqrt{L}$, then either there exists n such that $u_n(\beta) < 0$, or else $u_k(\beta)$ is zero for all $k > j$.*

Proof. We may assume that both $u_j(\beta)$ and $u_{j-1}(\beta)$ are nonnegative. For any positive integer N,

$$\begin{bmatrix} x-K & -L \\ 1 & 0 \end{bmatrix}^N \begin{pmatrix} u_j \\ u_{j-1} \end{pmatrix} = \begin{pmatrix} u_{N+j} \\ u_{N+j-1} \end{pmatrix}.$$

Diagonalizing the matrix evaluated at β, we find that the first row is

$$\frac{1}{\sin\theta}\begin{pmatrix} L^N \sin(N+1)\theta & L^N \sin N\theta \end{pmatrix},$$

where $\frac{1}{2}(\beta - K) = L\cos\theta$, and $\lambda = Le^{i\theta}$ is the eigenvalue with $0 < \theta < \pi$. Hence $u_{N+j}(\beta) = (L^N/\sin\theta)(u_j(\beta)\sin(N+1)\theta + u_{j-1}(\beta)\sin N\theta)$. Since $\sin\theta > 0$, and we may choose N so that both $\sin N\theta$ and $\sin(N+1)\theta$ are strictly negative, we see that

either both $u_{j-1}(\beta)$ and $u_j(\beta)$ are zero (which entails that $u_k(\beta) = 0$ as desired), or some $u_n(\beta) < 0$. □

LEMMA 2.2. *Let $\{c_n\}$ and $\{d_n\}$ be sequences of real numbers and let*

$$\{v_0 = 1, v_1 = x, \cdots, v_{i+1} = (x - c_i)v_i - d_i v_{i-1}, \cdots\}$$

be the sequence of polynomials obtained therefrom. Suppose that $a_i \geq c_n$ and $b_i \geq d_n$ for all $i, j \leq n$. If β is a real number such that $v_n(\beta) < 0$, there exists $m \leq n$ with $u_m(\beta) < 0$.

Proof. By [AI, (2.10)], we may write $v_t = \sum_{k=1}^{t} f(k,t)u_k$ with $f(k,t)$ nonnegative and real. Evaluating at $t = n$ and $x = \beta$ yields the result. □

LEMMA 2.3. *Let K and L be elements of $\mathbf{R} \cup \{\infty\}$ defined as $K = \sup a_i$ and $L = \sup b_j$, let ϵ be a positive real number. Suppose that β is a real number less than $K - \epsilon + 2(L - \epsilon)^{1/2}$. Assume there exists j such that for all $k \geq j$, $a_k \geq K - \epsilon$ and $b_k \geq L - \epsilon$ (if K or L is ∞, then the corresponding condition is vacuous). Then either there exists n such that $u_n(\beta) < 0$, or else $u_m(\beta) = 0$ for all $n > j$.*

Proof. Applying Lemmas 2.1 and 2.2 with K and L replaced by $K-\epsilon$, $L-\epsilon$, respectively, the first j values of c_n and d_n are $\{0, 0, 0, \cdots, 0, K - \epsilon\}$, $\{0, 0, 0, \cdots, 0, L - \epsilon\}$, respectively. □

LEMMA 2.4. *Suppose that β is a real number exceeding $K + 2L^{1/2}$ and $a_i = K$, $b_j = L$ for all i and j. Then $u_i(\beta) > 0$ for all i.*

Proof. Using the matrix formulation as in Lemma 2.1, we have that

$$u_N(\beta) = ((\beta - K)^2 - 4L)^{-1/2} \cdot ((\beta - \lambda^-)(\lambda^+)^{N+1} - (\beta - \lambda^+)(\lambda^-)^{N+1})$$

where $\lambda^{\pm} = \frac{1}{2}\left(\beta - K \pm ((\beta - K)^2 - 4L)^{1/2}\right)$.

Note that $\beta - \lambda^- > \beta - \lambda^+$ and $\lambda^+ > \lambda^- > 0$, so that $u_N(\beta) > 0$. (In fact, it is easy to verify that $u_N(\beta) > u_{N-1}(\beta)$.) □

LEMMA 2.5. *Suppose $a_i \leq c_n$ and $b_j \leq d_n$ for all $i, j \leq n$ and the polynomials $\{v_i\}$ are obtained recursively from the three-term difference equation corresponding to $\{c_i\}$ and $\{d_i\}$ in place of $\{a_i\}$ and $\{b_i\}$. Let β be a real number. If $v_i(\beta) > 0$ for all i, then $u_i(\beta) > 0$ for all i.*

Proof. This follows exactly as in Lemma 2.2. □

So if $M = K + 2L^{1/2}$ is finite and β exceeds M, then for all n, $u_n(\beta) > 0$. Thus $u_n(M) \geq 0$ (when M is finite). Regardless of whether M is finite, β less than M entails either $u_m(\beta) < 0$ for some m, or $u_n(\beta) \geq 0$ for all n, and if the latter holds, equality occurs for at least one value of n. This latter possibility can occur, and results in complications (viz. §5).

Let β be a nonnegative real number such that $u_i(\beta) \geq 0$ for all i, and $u_m(\beta) = 0$ for some m. Call β a *zero of positivity* of $\{u_i\}$.

THEOREM 2.6. *Let β be a zero of positivity of $\{u_i\}$, and let m be the smallest integer such that $u_m(\beta) = 0$. Then*

(i) $\beta \leq M$;

(ii) *For all $n \geq m$, $u_n(\beta) = 0$;*

(iii) $b_1 = b_2 = \cdots = b_m = 0$ *and* $\beta = a_{m-1}$ *($a_0 = 0$) if the sequence $\{b_i\}$ is monotone;*

(iv) *If both sequences $\{a_i\}$ and $\{b_i\}$ are monotone, then $a_1, \cdots, a_{m-1}$ are all zeros of positivity and $a_{m-1} > a_{m-2}$.*

Proof. From Lemmas 2.2–2.5, part (i) follows. If $m = 1$, then $\beta = 0$, and as $u_2 = (x - a_1)x - b_1$ is nonnegative at β, we must have that b_1 is zero, so that

$u_1(\beta) = u_2(\beta) = 0$, which yields (ii) in this case. If $m > 1$, write

$$0 = u_m(\beta) = (\beta - a_{m-1})u_{m-1}(\beta) - b_{m-1}u_{m-2}(\beta),$$
$$u_{m+1}(\beta) = (\beta - a_m)u_m(\beta) - b_m u_{m-1}(\beta).$$

The latter equation, in combination with $u_{m-1}(\beta) \neq 0$ and $u_{m+1}(\beta) = 0$, yields $b_m = 0$. As $u_m(\beta) = u_{m+1}(\beta) = 0$, (ii) follows from the recurrence relation.

We have just seen that $b_m = 0$, so $b_1 = b_2 = \cdots = b_m = 0$ if monotonicity holds. Thus $u_m(\beta) = (\beta - a_{m-1})u_{m-1}(\beta)$, from which $\beta = a_{m-1}$, verifying (iii).

Finally,

$$u_i = x\left(\prod_{j=1}^{i-1}(x - a_i)\right)$$

for $i \leq m$, so that each of $a_1, \cdots, a_{m-2}$ is a zero of consecutive u_i, and positivity then follows from monotonicity of $\{a_i\}$. To conclude, $u_{m-1}(a_{m-1})$ is not zero, so $a_{m-1} > a_{m-2}$. □

Monotonicity is not necessary for the positivity condition on $\{u_i\}$ to hold (i.e., that $u_j u_k$ is always a nonnegative combination of the u's). For example, if $a_1 = a_3 = a_5 = \cdots = 1$, and $a_2 = a_4 = a_6 = \cdots = 2$, and $b_i = 0$ for all i, then $u_j u_k$ is always a nonnegative combination of the u's—this follows from a brief calculation. On the other hand, when $b_i = 0$ for all i, and the positivity condition on the u's holds, then $a_i < a_{i-1}$ entails that a_{i-1} belong to $\{a_1, \cdots, a_{i-2}\}$; hence if the a's are distinct, the sequence they form must be monotonic! The remarks in this paragraph are results of work obtained jointly with my former colleague Angelo Mingarelli.

PROPOSITION 2.7. *If $b_1 = b_2 = \cdots = b_m = 0$ and $0 \leq a_1 \leq a_2 \leq \cdots \leq a_{m-1}$ and the positivity condition holds on the u's (e.g., if the sequences $\{a_i\}$ and $\{b_i\}$ are monotone), and $\{a_i\}$ is monotone, then each of $a_1, a_2, \cdots, a_{m-1}$ is a zero of positivity of $\{u_i\}$.*

Proof. An easy verification of the definitions. □

So Proposition 2.7 gives a prescription for zeros of positivity to exist, and Theorem 2.6 gives necessary conditions for their existence. The theorems we prove in §4 will hypothesize no such zeros. The general case is discussed in §5.

Let u be an arbitrary element of $\mathbf{R}[x]^+$. We say u is *gapless* if on writing $u = \sum_{i=0}^n \lambda_i u_i$ with $\lambda_n > 0$, all of the real numbers λ_i are strictly positive. We say an element u of $\mathbf{R}[x]^+$ is *solid* if some power of u is gapless.

LEMMA 2.8. *Under the assumption that $\{u_i\}$ satisfies the positivity condition on products, a product of gapless elements is gapless, and a product of solid elements is solid.*

Proof. Write $u = \sum_{i=0}^n \lambda_i u_i$ and $v = \sum_{j=0}^m \mu_j u_j$, with all of the coefficients strictly positive. Given k with $0 \leq k \leq m + n$, write $i + j = k$ for some i and j. Then the element $c = \lambda_i \mu_j u_i u_j$ appears in uv in the sense that $uv \geq c$, and $u_i u_j$ is a positive element of degree $i + j = k$. Thus u_k must appear in it, and so u_k appears (i.e., with nonzero coefficient) in uv. The second part is an obvious consequence of the definitions. □

If R is a partially ordered ring and $\alpha: R \to \mathbf{R}$ is a state, we say that α is *faithful* if $R^+ \cap \ker \alpha = (0)$.

PROPOSITION 2.9. *If the element u of $\mathbf{R}[x]^+$ is solid and is not constant, then*

(a) *u^{-1} belongs to R_u^+;*

(b) *$R_u[u] = S_u$; this means that the ring obtained by adjoining the inverse of u^{-1} (namely u) to R_u is S_u;*

(c) *A pure state of* $(R_u, 1)$ *extends to a (multiplicative) state on* S_u *if and only if its value at* u^{-1} *is not zero;*

(d) *The pure states of* $(R_u, 1)$ *are of the following kinds:*

(i) *Point evaluation at a zero of positivity of the* $\{u_i\}$*;*

(ii) *Point evaluation at a point in the interval* (M, ∞) *if* $M < \infty$*;*

(iii) $f \mapsto \lim_{t\to\infty} f(t)$ *(for* f *in* R_u*);*

(iv) *Point evaluation at* M *if* $M < \infty$.

States of type (i) *or* (iii) *are not faithful, those of type* (ii) *are faithful, and the state of type* (iv) *is faithful if and only if* M *is not a zero of positivity of* $\{u_i\}$.

Proof. (a) By solidity, there exists an integer k so that u^k is gapless. Hence there exists an integer N with $u^{k-1} \leq Nu^k$ in $\mathbf{R}[x]$. Hence $u^{-1} \leq N$ in S_u and u^{-1} obviously belongs to S_u^+, so u^{-1} lies in R_u^+.

(b) By definition, $S_u = \mathbf{R}[x, u^{-1}]$; now $R_u[u] = R_u[(u^{-1})^{-1}]$. As $R_u \subset S_u$, it follows that $R_u[u] = S_u$. It suffices to show that x belongs to $R_u[u]$, for which xu^{-k} (the same k as in the proof of (a)) belonging in R_u is sufficient. However, $x = u_1 \leq Ku^k$ for some integer K, the inequality occurring in $\mathbf{R}[x]$. Thus xu^{-k} belongs to R_u.

(c) Let ϕ be a faithful pure state of R_u. Then $\phi(u^{-1}) \neq 0$, so ϕ extends to a real-valued map on $R_u[u] = S_u$ via $\phi(ru^m) = \phi(r)/\phi(u^{-m})$. As in [H1, p. 7], it is easily verified that this is a multiplicative state of S_u, and moreover, the full assumption of faithfulness can be replaced by $\phi(u^{-1}) \neq 0$. Conversely, if ϕ is a multiplicative state of S_u, then its restriction to R_u is a multiplicative (hence pure) state of R_u; moreover, the extension (from R_u to S_u) described above will restrict to the original state.

(d) Certainly, each of the four types of function $R_u \to \mathbf{R}$ is a multiplicative, hence pure, state of R_u. Let $\phi: R_u \to \mathbf{R}$ be a pure state; it is multiplicative (Theorem 1.1). Suppose to begin with that $\phi(u^{-1}) \neq 0$. Then ϕ extends to a multiplicative state on $S_u = \mathbf{R}[x, u^{-1}]$, so is given as a point evaluation of x at some point b in $\mathbf{R}^+$. By Lemmas 2.2–2.5, either $b \leq M$ and b is a zero of positivity (i), or $b > M$ (and $M < \infty$) and the state is faithful (ii), or $b = M$. The final statement, concerning the case $b = M$, is straightforward.

This leaves the possibility that $\phi(u^{-1}) = 0$. Suppose $n = \deg u$. For $i \leq n$, observe that $u_i u^k \leq Ku^{k+1}$ (as u^{k+1} is gapless) for some integer K, so u_i/u belongs to R_u. If $i > n$, we similarly have u_i^n/u^{n-1} in R_u. Thus when $i < n$,

$$(*) \qquad \left(\frac{u_i}{u}\right)^n = \left(\frac{1}{u}\right) \cdot \left(\frac{u_i^n}{u^{n-1}}\right) \in \frac{1}{u} \cdot R_u.$$

Evaluate $(*)$ at the state ϕ; as ϕ is multiplicative and $\phi(u^{-1}) = 0$, we deduce that $\phi(u_i/u) = 0$ for $i < n$. Set $\phi_0(f) = \lim_{t\to\infty} f(t)$ (for f in R_u). Now the quotient of R_u by the principal ideal generated by u^{-1}, $R_u/(u^{-1})$, is spanned by the images of $\{u_i/u^{-1} \mid i = 1, 2, \cdots, n\}$, so that the kernel of ϕ equals that of ϕ_0. As $\phi(1) = \phi_0(1) = 1$, we have that $\phi = \phi_0$. □

3. Algebraic properties of R_u. Having described the pure states of R_u, we now work out some algebraic properties necessary to prove the main results. We continue the notation from §2.

LEMMA 3.1. *If* u *is solid, then* R_u *is integrally closed in its field of fractions.*

Proof. Select au^{-m} in the integral closure of R_u in S_u; the latter is $R_u[u] = \mathbf{R}[x, u^{-1}]$, so is integrally closed. Since au^{-m} satisfies a monic polynomial with coef-

ficients from R_u, it therefore is bounded at ∞. Hence $\deg a \leq \deg u_m = m \deg u$. As u^k is gapless for all sufficiently large k, we have that $-Ku^k \leq au^{k-m} \leq Ku^k$ for some positive integer K. Thus au^{-m} belongs to R_u. □

LEMMA 3.2. *If u is solid, then R_u is a finitely generated* **R***-algebra, hence is noetherian.*

Proof. Let n denote the degree of u. We will show that $R_u = \mathbf{R}[x^j/u; j \leq n]$. Obviously x^j/u belongs to R_u if $j \leq n$ (the argument in Lemma 3.1 works, for example). Now R_u is spanned as a real vector space by

$$\{u_i u^{-m} \mid i \leq m\}$$

and thus is spanned by $\{x^i u^{-m} \mid i \leq mn\}$. For each pair i, m with $i \leq mn$, write $i = cn + j$, where c and j are nonnegative integers and $0 \leq j < n$; then

$$\frac{x^i}{u^m} = \left(\frac{x^n}{u}\right)^c \cdot \frac{x^j}{u} \cdot \left(\frac{1}{u}\right)^{m-c-1}.$$

Thus $x^i u^{-m}$ belongs to $\mathbf{R}[u_j/u; j \leq n]$. □

Even without solidity, the conclusion of Lemma 3.2 is true (although that of Lemma 3.1 is not); the proof is more complicated. Lemmas 3.1 and 3.2 yield that R_u is a Dedekind domain whenever u is solid. Indeed, R_u is an order in $\mathbf{R}[x, u^{-1}]$, so it has Krull dimension one and is also an integrally closed noetherian ring. We now analyze the order ideals of R_u.

In a partially ordered abelian group G, an *order ideal* is a subgroup H such that $H = H \cap G^+ - H \cap G^+$, and for $g \in G$, $0 \leq g \leq h \in H$ implies that g belongs to H. By [H1, §I], order ideals in a partially ordered ring having 1 as an order unit are ideals. We first determine which order ideals are prime (as ideals).

LEMMA 3.3. *If u is solid, then the order ideals of R_u which are prime as ideals of R_u are of the form:*

(i) *The kernel of the state $f \mapsto f(\beta)$ where β is a zero of positivity of $\{u_i\}$ (however, such an ideal need not be an order ideal!); or*

(iii) *The kernel of the state $\phi_0 : f \mapsto \lim_{t\to\infty} f(t)$; this is, $\{au^{-m} \mid \deg a < m \deg u\}$. The kernel of ϕ_0 is a prime order ideal.*

Proof. Let I be a nonzero prime ideal of R_u that is also an order ideal. As the Krull dimension of R_u is 1, R_u/I is a field. As R_u is a finitely generated real algebra, R_u/I is algebra isomorphic to one of $\mathbf{R}$ or $\mathbf{C}$. However, R_u/I inherits the quotient ordering; it is thus partially ordered (and directed), so in fact $R_u/I = \mathbf{R}$. As the quotient map $R_u \to R_u/I$ is positive, I is the kernel of a pure state, and of course I is generated (additively) by $I \cap R_u^+$. As states of type (ii) (Proposition 2.9) are faithful, this type is excluded. This leaves types (i) and (iii).

In the type (iii) situation, $\{au^{-m} \mid \deg a \leq m \deg u\}$ is spanned by $\{u_j u^{-m} \mid j \leq m \deg u\}$, so the former is an order ideal. □

Suppose that $a_1 < a_2$ and $b_1 = b_2 = 0$ (so that a_0 and a_1 are zeros of positivity of $\{u_i\}$). We may define pure states of R_u, α_0 and α_1, via $\alpha_i(f) = f(a_i)$. If u is solid, these two states are pure, but

$$(0) \neq \ker \alpha_0 \cap R_u^+ \subsetneq \ker \alpha_1 \cap R_u^+.$$

This implies that the kernel of α_1 is not an order ideal. Moreover, $\ker \alpha_0$ is an order ideal, but its square is not.

PROPOSITION 3.4. *If u is solid and $\{u_i\}$ admits no zeros of positivity, then every order ideal of R_u is of the form I_∞^k ($k = 1, 2, \cdots$), where $I_\infty = \ker \phi_0$. Conversely, I_∞^k is always an order ideal.*

Proof. Let I be an order ideal of R_u. Then R_u/I is a partially ordered ring with order unit, so admits a state ϕ; hence there exists a pure state ϕ of R_u such that $\phi(I) = 0$. As $I^+ \subset \ker \phi$ and states of type (i) are excluded, we are left with $\phi = \phi_0$, and so $I \subseteq I_\infty$. Suppose that I is contained in some other maximal ideal $\mathcal{M}$ (that is not necessarily an order ideal).

As I is generated by its positive cone, I^+ is generated as an $\mathbf{R}^+$-semigroup by certain elements of the form $r = u_i u^{-m}$ with $i \leq \deg u^m = mn$. Then $r^n = u_i^n/u^{mn} = (1/u^s)(u_i^n/u^i)$ where $s = mn - i$. As $\deg u_i^n = \deg u^i$, u_i^n/u^i belongs to R_u and thus to R_u^+. Now if γ is any pure state of R_u, $\gamma(u_i^n/u^i) > 0$ by the hypotheses and Proposition 2.9(d), so that u_i^n/u^i is an order unit of R_u (Theorem 1.1). Hence there exists a positive integer N such that $1 \leq N u_i^n/u^i$ in R_u. As I is convex and $0 \leq 1/u^s \leq N r^m$, we have that u^{-s} belongs to I. As $\mathcal{M}$ is maximal, $1/u$ belongs to it.

Now suppose that $j < mn$ and consider $(u_j/u^m)^n = (1/u)^l (u_j^n/u^j)$ for $l = mn - j$. As above, we deduce that u_j/u^m belongs to $\mathcal{M}$, so that $I_\infty \subseteq \mathcal{M}$; thus equality holds. Hence I_∞ is the only maximal ideal in R_u containing I. As R_u is a Dedekind domain, $I = I_\infty^k$ for some integer k.

Now we show that I_∞^k is always an order ideal (it is not generally true that powers of an order ideal are order ideals). Clearly

$$I_\infty^k \subseteq \{ a u^{-m} \mid \deg a \leq mn - k \} .$$

First we show that $x^{n-j}/u \in I_\infty^j$ for $j < n$. We note that u^{j-1} is a polynomial of degree $(j-1)n$. Every monomial in $x^{n-j}u^{j-1}$ is of degree at most $(n-1)j$. Let x^d be a monomial appearing therein. Then $d \leq (n-1)j$, so we may write $d = \sum_{i=1}^j d(i)$ for positive integers $d(i) \leq n-1$. Then $x^d/u^j = \prod x^{d(i)}/u \in I_\infty^j$. So $x^{n-j}/u = x^{n-j}u^{j-1}/u^j$ is a sum of elements of I_∞^j. It similarly follows (on replacing u by u^m) that $x^{mn-k}/u^m \in I_\infty^k$. The inclusion reverse to the displayed one readily follows.

Since $\{u_i\}_{i=0}^n$ is a basis for the set of polynomial of degree n or less, we obtain that I_∞^k is spanned by $\{ u_i u^{-m} \mid i \leq mn - k \}$. Thus the ideal is directed. Finally if s is an element of I_∞^k and $0 \leq r = a'/u^{m'} \leq s = a/u^m$, it is immediate that $\deg a' \leq m'n - k$, so $r \in I_\infty^k$. Thus I_∞^k is convex and directed, so is an order ideal. □

(Part of the reason for the length of this argument is that R_u is not necessarily a principal ideal domain—at worst its class group is of order 2, generated by the class of I_∞.)

An ideal J of a (commutative) ring is called *primary* if, for s and t in the ring, whenever the product st belongs to J and neither s nor t lies in J, then there exists an integer m so that s^m and t^m both belong to J. In Dedekind domains, powers of prime ideals are primary. The outcome of the preceding is that if u is solid and $\{u_i\}$ admits no zeros of positivity, then every order ideal of R_u is primary. This can fail if $\{u_i\}$ admits zeros of positivity.

4. Positivity results (without zeros of positivity).

THEOREM 4.1. *Let $\{u_0 = 1, u_1 = x, \cdots, u_{n+1} = (u_n - a_n)u_n - b_n u_{n-1}, \cdots\}$, be a sequence of polynomials such that*

(a) *$\{u_i\}$ satisfies the positivity condition.*

(b) *On defining* $M = \sup a_i + 2\sup\sqrt{b_i}$ *in* $\mathbf{R} \cup \{\infty\}$, *if* $M < \infty$, *then* $M = \lim a_i + 2\lim\sqrt{b_i}$.

(c) $\{u_i\}$ *has no zeros of positivity.*

(*For example, if* $\{a_i\}$ *and* $\{b_i\}$ *are monotone nondecreasing, then conditions* (a) *and* (b) *hold.*) *Let* f *and* u *be elements of* $\mathbf{R}[x]$, *with* u *positive and solid with respect to* $\{u_i\}$. *There exists* N *so that* $u^N f$ *is positive if and only if*

When $M < \infty$, f *is strictly positive on the interval* $[M, \infty)$

When $M = \infty$, *the leading coefficient of* f (*in terms of* x) *is positive.*

Proof. We may assume (replacing u by a suitable power of itself) that u is already gapless, and so are all of its powers. Form R_u. We have previously determined (Proposition 2.9) that its states are given by $[M, \infty]$, where the state at ∞ is ϕ_0. Select k so that $m = \deg f \leq k \deg u$. As u^k is gapless, we have that $r = fu^{-k}$ belongs to R_u.

Set $v = u_0 + u_1 = 1 + x$ and form R_v; obviously $s = f/v^m$ lies in R_u. We shall show that s belongs to R_v^+ and then use this to prove that r belongs to R_u^+, which is what we want. As $\deg f = \deg v^m$, $\lim_{t\to\infty} s(t) \neq 0$, so by hypothesis (either case) $\phi_0(s)$ exceeds zero. If $M < \infty$, then $s(t) > 0$ for all $t \geq M$ by hypothesis. Thus s is strictly positive at all pure states of R_v, so in particular by Theorem 1.1, s belongs to R_v^+. Hence there exists an integer e so that $(1+x)^e f$ is positive with respect to $\{u_i\}$. If $\deg u = n$, then z defined as $(1+x)^n/u$ is an order unit of R_u and rz^e belongs to R_u^+. If r does not belong to I_∞, then its value at ∞ is positive. The hypotheses guarantee that r is strictly positive at all pure states; by Theorem 1.1, r is already in R_u^+. So we need only consider the case that r belongs to I_∞.

Since rz^e is positive in R_u, we may consider the order ideal J that it generates (as an order ideal). By Proposition 3.4, $J = I_\infty^k$ for some integer k. Set $I = (r)$, the principal ideal generated by r. As $I_\infty^k \subseteq I \subseteq I_\infty$ and R_u is a Dedekind domain, we have that $I = I_\infty^j$ for some $j \leq k$, and this implies that I is also an order ideal. As I is finitely generated (as an ideal), I admits an order unit (considering I itself as a partially ordered group), y. We may write $y = ra$ for some a in R_u. Let γ be a pure state of (I, y). By [H1, 1.3], there exists a pure state L on R_u such that $1 = \gamma(y) = \gamma(r)L(a)$.

If $M < \infty$, then any point in $[M, \infty)$ yields a state such that $\alpha(y) > 0$, so that $\alpha/\alpha(y)$ is a pure state of (I, y); thus $\alpha(a) > 0$ for all points in $[M, \infty)$. If a vanishes at infinity, then y belongs to I_∞^{j+1}, which is itself an order ideal, so that y could not be an order unit for I_∞^j. Thus a is strictly positive at all the pure states of R_u, so $L(a) > 0$ for all such L. Hence $f(r) > 0$, so r is positive as an element of (I, y). Thus it lies in R_u^+.

If $M = \infty$, the same argument yields that $L(a) \neq 0$ if $L = \phi_0$; as this is the only state of R_u, we have that $\gamma(a)$ is constant (independent of f) and equals $1/L(a)$. However, $rz^e \geq 0$ entails $\gamma(r) > 0$ (as z^e is an order unit of R_u). Hence r belongs to I^+, and so to R_u^+. The converse is completely straightforward. □

Even if we allow zeros of positivity, these conditions on f are sufficient for eventual positivity, but not necessary. The same proof works; however, we will obtain complete results in this case in the next section.

Now we consider the condition $a_1 b_1 > 0$, equivalent to condition (A) of the Introduction. We show that in the presence of monotonicity, every nonconstant positive element of $\mathbf{R}[x]$ (with respect to $\{u_i\}$) is solid when this condition applies.

If u is an element of $\mathbf{R}[x]$, we say u *contains* u_j if the coefficient of u_j in the expansion of u with respect to $\{u_i\}$ is positive. If u is positive, then u contains u_j if and only if there exists a positive integer N such that $Nu \geq u_j$ (with respect to $\{u_i\}$). If u and v are both positive elements in $\mathbf{R}[x]$, we say u contains v if there exists a positive integer N with $Nu \leq v$.

PROPOSITION 4.2. *If $\{a_i\}$ and $\{b_i\}$ are monotone nondecreasing sequences of nonnegative numbers, then for $1 \leq j \leq k$, $u_j u_k - u_{j-1}u_{k+1} \geq 0$; if additionally $a_1 b_1 > 0$, then each of u_{k-j}, u_{k-j+1} is contained in $u_j u_k - u_{j-1}u_{k+1}$.*

Proof. These are proved via induction on j. If $j = 1$, we obtain

$$u_1 u_k = u_{k+1} + a_k u_k + b_k u_{k+j-2} \geq u_0 u_{k+1} = u_{k+1}.$$

In this case, both parts are proved.

For $j > 1$, we have

$$\begin{aligned} u_{j+1}u_{m+1} &= ((u_1 - a_j)u_j - b_j u_{j-1})u_{m+1} \\ &= u_j(u_1 - a_{m+1})u_{m+1} + (a_{m+1} - a_j)u_j u_{m+1} - b_j u_{j-1}u_{m+1} \\ &= u_j(u_{m+2} - b_{m+1}u_m) + (a_{m+1} - a_j)u_j u_{m+1} - b_j u_{j-1}u_{m+1} \\ &= u_j u_{m+2} + (a_{m+1} - a_j)u_j u_{m+1} + b_{m+1}(u_j u_m - u_{j-1}u_{m+1}) \\ &\quad + (b_{m+1} - b_j)u_{j-1}u_{m+1}. \end{aligned}$$

As the a's and b's are increasing, we obtain $u_{j+1}u_{m+1} - u_j u_{m+2} \geq 0$ by induction. If a_1 and b_1 are not zero, by induction, we deduce that u_{m-j} and u_{m-j+1} are contained in $u_j u_m - u_{j-1}u_{m+1}$; and since all the coefficients are positive, u_{m-j} and u_{m-j+1} are contained in $u_{j+1}u_{m+1} - u_j u_{m+2}$; but $m - j = m + 1 - (j + 1)$, and we are done. □

In at least one sense, the preceding is best possible; if $a_i = K$ and $b_i = L$ for all $i > 0$, then $u_j u_k - u_{j-1}u_{k+1} = L^j(Ku_{k-j+1} + Lu_{k-j})$ when $1 \leq j \leq k$.

LEMMA 4.3. *If $\{a_i\}$ and $\{b_i\}$ are monotone nondecreasing nonnegative sequences with $a_1 b_1 > 0$, then for $1 \leq i \leq k$, $u_j u_k$ contains u_m for all m satisfying $k - j \leq m \leq k + j$. In particular, u_k^2 contains every u_m for $0 \leq m \leq 2k$.*

Proof. By the first part of Proposition 4.2, we have that

$$u_j u_k \geq u_{j-1}u_{k+1} \geq \cdots \geq u_{j-i}u_{k+i} \geq \cdots \geq u_1 u_{k+j-1}.$$

By the second part, we obtain that $u_{j-i}u_{k+i}$ contains u_{k-j+2i} and $u_{k-j+2i+1}$ for $0 \leq i \leq j - 1$; that is, $u_{k-j}, u_{k-j+1}, \cdots, u_{k+i-1}$ appear in $u_j u_k$. Finally,

$$u_1 u_{k+j-1} = u_{k+i} + a_{k+j-1}u_{k+j-1} + b_{k+j-1}u_{k+j-2},$$

so that u_{k+j} appears in $u_j u_k$ (which is obvious anyway). □

LEMMA 4.4. *Under the preceding hypotheses, if u in $\mathbf{R}[x]^+$ is nonconstant, then u^2 is gapless.*

Proof. Writing $u = \alpha_k u_k + \sum_{j<k} \alpha_j u_j$, with $\alpha_k > 0$ and $\alpha_j \geq 0$, we see that u^2 contains u_k^2, and the result follows from Proposition 4.2. □

THEOREM 4.5. *If $\{a_i\}$ and $\{b_i\}$ are monotone sequences with $a_1 b_1 > 0$, then all nonconstant u in $\mathbf{R}[x]^+$ are solid, and the conclusion of Theorem* 4.1 applies.

5. Positivity results (allowing zeros of positivity). This section is considerably more technical than its predecessors and can safely be omitted on a first reading of the article. Throughout, we assume that the sequences $\{a_i\}$ and $\{b_i\}$ are monotone (nondecreasing). We allow $\{u_i\}$ to have zeros of positivity. This means that if $b_n = 0$ for some n, then $0\ (= a_0), a_1, \cdots, a_n$, are all zeros of positivity by Theorem 2.6 and

Proposition 2.7. If β is a zero of positivity of $\{u_i\}$, we say it has *multiplicity* k (with k in $\mathbf{N} \cup \{\infty\}$) if $k = \#\{i \mid a_i = \beta \text{ and } b_{i+1} = 0\}$. Note that if the multiplicity of β is infinite, then all of the b's are zero, and for all sufficiently large i, $a_i = \beta$ (so that M would equal β as well).

Fix an element u which is positive, solid, and of degree n. Let I be an order ideal of R_u that is not contained in I_∞. We describe all possibilities for such I, showing that they form a chain of ideals, each of codimension 1 in the next.

The first step is to show that the principal ideal of R_u, $(1/u)$, is an order ideal thereof. If $0 \leq r \leq 1/u$ in R_u, upon writing $r = fu^{-k}$, with f in $\mathbf{R}[x]^+$, there exists n that $0 \leq fu^{n-k} \leq u^{n-1}$ in $\mathbf{R}[x]$. Thus $t = fu^{-(k-1)}$ belongs to R_u, and $r = (1/u)t$ belongs to $(1/u)$. Thus $(1/u)$ is convex. As $1/u$ is clearly an order unit for $(1/u)$ (as a partially ordered subgroup of R_u) the latter is directed, and so is an order ideal.

Next, we show that the only maximal ideal containing $1/u$ is I_∞. Let $\mathcal{M}$ be another maximal ideal containing $1/u$. Then the third paragraph of the proof of Proposition 3.4 is valid (since it follows from the solidity of u that u_j^n/u^j belongs to R_u), so $I_\infty \subseteq \mathcal{M}$, a contradiction. As $I + I_\infty = R_u$, it follows that $I + (1/u) = R_u$.

Let $J = IS_u$; it is easily checked that this is an order ideal of S_u (in addition to being an ideal). Let i be the smallest integer such that u_i belongs to J; note that $J = J^+ - J^+$ so such an i exists. Of course, $i > 0$, as u_0 is the constant function 1. From a simple degree argument, it follows that $u_i u_j \geq K u_{i+j}$ in $\mathbf{R}[x]^+$. We deduce that u_{i+j} belongs to J for all $j \geq 0$.

Next, $u_i u_1 = u_{i+1} + b_i u_{i-1} + a_i u_i$, so if b_i were not zero, we would obtain $u_{i-1} \leq (1/b_i)u_i$, and thus u_{i-1} would belong to J, a contradiction. Hence $b_i = 0$, so that $b_m = 0$ for $m \leq i$.

Let c be a positive element of J; there exists k so that a defined as $u^k c$ is in $\mathbf{R}[x]^+$ (with respect to $\{u_i\}$). Write $a = \sum \lambda_j u_j$ with nonnegative real numbers λ_j. If λ_j exceeds zero, we obtain that the corresponding u_j belongs to J. Hence $\lambda_j > 0$ only for $j \geq i$.

For each $j \geq i$, define

$$d(j) = \begin{cases} \frac{j}{n} & \text{if } n \text{ divides } j \\ [\frac{j}{n}] + 1 & \text{else.} \end{cases}$$

We also define $e_j = u_j/u^{d(j)}$. Now we show that $I = \sum_{j \geq i} e_j \cdot R_u$; that is, I is generated as an R_u-ideal by an appropriate set of e_j's. Set I' to be the latter; this is an ideal, and clearly I' is not contained in I_∞ and $I' \cdot S_u = J$. Given $j \geq i$, there exists k so that u_j/u^k belongs to I; we must have that $j \leq kn$ (as u_j/u^k belongs to R_u), so $k \geq d(j)$. Hence $e_j(1/u)^{k-d(j)}$ belongs to I. As $I + (1/u)R_u = R_u$, it follows that e_j belongs to I. Thus $I' \subseteq I$.

To show $I \subseteq I'$, it suffices to show that $I^+ \subset I'$. To this end, pick c in I^+; we may write $c = \sum \lambda_j u_j/u^a$, where the λ's are nonnegative real numbers, and $0 \leq j \leq na$. Hence for each such j, $a \geq d(j)$. Thus c belongs to $\sum e_j \cdot R_u$, as desired.

Define the ideal $I_i = \sum_{j \geq i} e_j \cdot R_u$ for each i such that $b_i = 0$. We have just seen that every order ideal not contained in I_∞ is of the form I_i for some such i. Now we show that this I_i is always an order ideal.

Clearly $I_i S_u = \sum_{j \geq i} e_j S_u$, and this is an order ideal of S_u (as $b_0 = b_1 = \cdots = b_i = 0$), so $I_i = I_i S_u \cap R_u$ is an order ideal of R_u. (Recall that R_u is an order ideal of S_u, although not an ideal thereof; as S_u admits the Riesz interpolation property, finite intersections of order ideals are order ideals.) Obviously $I_i \subset I_i S_u \cap R_u$, but

essentially the same arguments as those given just above will show that equality holds. So I_i is an order ideal.

We have $R_u = I_0 \supset I_1 \supset \cdots$ (strict inclusion, continuing until the corresponding b_n is not zero, or forever if all of the b's are zero). Since $(1/u) + I_i = R_u$, we have that if $i \leq j$, then $I_i/I_j = (I_i S_u)/(I_j S_u)$, and the latter is of course a vector space with basis $\{u_i + I_j S_u, \cdots, u_{j-1} + I_j S_u\}$, hence has dimension $j - i$.

Since R_u is a noetherian ring, every ideal is finitely generated. It follows that every order ideal contains a relative order unit (that is, it admits an order unit on being viewed as a partially ordered group). We determine the pure states of I_i.

By [H1, I.1], all pure states, γ, on I_i are either (renormalized) restrictions of pure states of R_u, or have the following special form: There exists a pure state of R_u, L, such that $L(I_i) = 0$ and $\gamma(ar) = \gamma(a)L(r)$ for a in I_i and r in R_u.

Let γ be a pure state of I_i. If γ is a renormalized restriction of a state of R_u, then it cannot be evaluation at any a_j with $j < i$ (as these kill I_i). Provided that $a_i > a_{i-1}$, evaluation at a_i is a nonzero state. All other pure states of R_u restrict to (nonzero) states of I_i.

If γ is of the second kind, satisfying $L(r)\gamma(a) = \gamma(ar)$ and $L(I_i) = 0$, then the only candidates for L are evaluations at a_j with $j < i$ (and $b_i = 0$), since these are the only pure states that annihilate I_i. We show in fact that the only one that can arise is evaluation at the specific a_i, and *not* evaluation at any a_j strictly less than a_i, and in this case $a_i = a_{i-1}$ (of course, some of the a's may equal a_i and so yield the same L).

If I_{i+1} does not exist (that is, $b_{i+1} \neq 0$), then the smallest order ideal containing I_i^2 is I_i itself, and so I_i admits *no* such states at all. If I_{i+1} exists, then there exists $j \geq i$ such that the order ideal generated by I_i^2 is I_j. All such states (with their corresponding L's) induce states on I_i/I_j; this ordered vector space is order isomorphic to $I_i S_u / I_j S_u$. An easy computation yields that $u_i + I_i S_u$, is an order unit for the latter. Now γ extends to a state on $I_i S_u$ (via $\gamma(cu^k) = \gamma(c)L(u^{-1})^k$); we normalize so that $\gamma(u_i) = 1$.

From $u_i u_1 = u_{i+1} + a_i u_i$ (recalling that $b_i = 0$), we deduce that $L(u_1) = \gamma(u_i u_1) = \gamma(u_{i+1}) + a_i$. Since $u_1 = x$ and L is evaluation at some positive zero β (which is less than or equal to a_i), we deduce $\beta = \gamma(u_{i+1}) + a_i$. This yields $\beta = a_i$ and $\gamma(u_{i+1}) = 0$. It similarly follows that $\gamma(u_{i+k}) = 0$ for all $k > 0$. Hence there is at most one such "extra" state on I_i. Moreover, if $a_i \neq a_{i+1}$, then I_i^2 is not contained in I_{i+1} (as $I_{i+1} S_u = I_i(x - a_i)S_u$, and $x - a_i$ is relatively prime to $I_i S_u$, hence to its square). So again there are actually no other choices for γ except restrictions of states of R_u.

To summarize, we have deduced the following: The pure states of I_i are all given by:

(a) Restrictions of pure states of R_u, except for evaluations at zeros of positivity $\beta \leq a_{i-1}$;

(b) (After extension to $I_i S_u$) $u_{i+k} \mapsto 0$ for $k > 0$; this yields a state only if $a_i = a_{i-1}$ and $b_{i+1} = 0$.

It follows that I_i/I_j is totally ordered for $j > i$.

Now we can state and prove the main theorem of this section.

THEOREM 5.1. *Let $\{a_i\}$ and $\{b_i\}$ be monotone nondecreasing sequences of nonnegative real numbers, and let u be solid with respect to $\{u_i\}$. Let f be an element of $\mathbf{R}[x]$. There exists an integer N such that $u^N f \in \mathbf{R}[x]^+$ if and only if each of the following conditions hold:*

(1) *When M is finite, then $f|(M, \infty)$ is strictly positive; when $M = \infty$, the leading coefficient of f is positive.*

(2) *When M is finite and not a zero of positivity of the sequence $\{u_i\}$, $f(M) > 0$.*
(3) *For all zeros of positivity β of $\{u_i\}$, $f(\beta) \geq 0$.*
(4) *If $\beta < \beta'$ and both are zeros of positivity with $f(\beta') = 0$, then $f(\beta) = 0$ and the order of β as a zero of f must be at least as large as the multiplicity of β with respect to $\{u_i\}$.*
(5) *If $f = \rho_i u_i + \sum_{j>0} \rho_{i+j} u_{i+j}$, where the ρ's are real numbers and ρ_i is not zero, and if $b_{i+1} = 0$ and $a_i = a_{i-1}$, then $\rho_i > 0$.*

Proof. Necessity of (1) through (4) is easy to verify since u is strictly positive at all pure states of S_u; (5) follows from the observation that under the conditions imposed, the sign of the coefficient of u_i in $u^n f$ will always be that of ρ_i.

The proof of sufficiency will parallel that in Theorem 4.1. Let $k = \deg f$, and set $v = 1 + x = 1 + u_1$; then $s = f/v^k$ belongs to R_v, but not to I_∞ (as computed with respect to R_v). If s does not belong to I_1 (or I_1 does not exist), then s is strictly positive at all pure states of R_v by (1), (2), (3). So s belongs to R_v^+.

Otherwise, there exists j such that s belongs to I_j and either I_{j+1} does not exist, or s does not belong to I_{j+1}. The pure states of I_j have been computed above, and we evaluate s at them. Clearly (1) and (2) imply that s is strictly positive at points other than zeros of positivity, and s not in I_∞ yields that $\phi_0(s) \neq 0$. Hence (1) implies $\phi_0(s) > 0$. If $s(a_k) = 0$ for some positive zero (which must necessarily be of the form a_k, and at the same time, $b_{k+1} = 0$) with $a_k > a_j$, hypothesis (4) would yield that s belongs to I_{j+1}. Hence s is strictly positive at all of the nonzero restrictions of the pure states of R_v to I_j.

There remains the case that $a_j = a_{j-1}$ and $b_{j+1} = 0$, and the state that induces the map $I_j S_v \to \mathbf{R}$ be given by (after a suitable renormalization) $u_{j+k} \mapsto 0$ for $k > 0$. The corresponding map applied to f (as an element of $I_j S_v$!) sends it to ρ_j. It must be shown that j equals the i appearing in (5).

Write $f = \sum \rho_t u_t$ in $\mathbf{R}[x]$. Clearly $\rho_t = 0$ for $t < j$. If $\rho_t = 0$, then following from $a_j = a_{j-1}$, we would deduce that s belongs to I_{j+1}, and by hypothesis (5), $\rho_j > 0$. Thus s is strictly positive at all of the pure states of I_j, and so s belongs to I_j^+.

So in all cases, there exists an integer m such that $(1 + x)^m f$ belongs to $\mathbf{R}[x]^+$ (computed with respect to the basis $\{u_i\}$).

Recalling $k = \deg f$ and $n = \deg u$, select an integer p so that $k \leq pn$, and set $r = f/u^p$; this is an element of R_u. Setting $z = (1 + x)^{mn}/u^m$, we see that z is an order unit of R_u, and moreover $rz \in R_u^+$. Let J be the order ideal generated by the positive element rz, i.e.,

$$J = \{\, s \in R_u \mid \text{there exists a positive integer } N \text{ with } -Nrz \leq s \leq Nrz \,\}.$$

This is also an ideal of R_u, and we will show that it contains r. First note that rz belongs to I_∞^k if and only if r belongs to I_∞^k, since $\phi_0(z) > 0$. Next, if rz belongs to I_i for some i, we have that $(1 + x)^{mn} f$ belongs to $\sum_{j \geq i} u_j S_u$, and as $1 + x$ is relatively prime to each of the u's (within the ring S_u), we deduce that f belongs to $\sum_{j \geq i} u_j S_u$ as well. Thus r belongs to $\sum_{j \geq i} u_j S_u \cap R_u = I_i S_u \cap R_u$, and this is nothing but I_i.

Since R_u is Dedekind, every order ideal of R_u is of the form $I_i \cap I_\infty^k$ or I_∞^k for suitable i and k, and it follows that r belongs to J. Let γ be a pure state of J. Then there exists a pure state L of R_u such that $L(rz) = \gamma(r)L(z)$ (either γ is the restriction after renormalization of L, or else $L(J) = 0$ [H1, 1.3]). As z is an order unit of R_u, $L(z) > 0$; as rz is an order unit of J, $\gamma(rz) > 0$. We thereby deduce that $\gamma(r) > 0$. Hence r belongs to J^+, and hence to R_u^+, as desired. □

Acknowledgments. I would like to thank Norbert Riedel for suggesting that the methods of [H1] and related ideas could be applied to the type of questions discussed here, and Angelo Mingarelli for some interesting conversations about the material.

REFERENCES

[AI] R. ASKEY AND M. ISMAIL, *Recurrence relations, continued fractions and orthogonal polynomials*, Mem. Amer. Math. Soc., 300 (1984), p. 108.

[H1] D. HANDELMAN, *Positive polynomials and product type actions of compact groups*, Mem. Amer. Math. Soc., 320 (1985), p. 79.

[H2] ———, *Deciding eventual positivity of polynomials*, Ergodic Theory and Dynamical Systems, 6 (1986), pp. 57–79.

[Po] H. POINCARÉ, *Sur les équations algébriques*, C.R. Acad. Sciences, 97 (1883), pp. 1418–1419.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 1065-1082, July 1990
© 1990 Society for Industrial and Applied Mathematics
014

L_∞ MARKOV AND BERNSTEIN INEQUALITIES FOR FREUD WEIGHTS*

A. L. LEVIN† AND D. S. LUBINSKY‡

Abstract. Let $W(x) := e^{-Q(x)}$, $x \in \mathbb{R}$, where $Q(x)$ is even and continuous in $\mathbb{R}$, $Q(0) = 0$, and Q'' is continuous in $(0, \infty)$ with $Q'(x) > 0$ in $(0, \infty)$, and for some $C_1, C_2 > 0$,

$$C_1 \leqq (xQ'(x))'/Q'(x) \leqq C_2, \qquad x \in (0, \infty).$$

For example, $Q(x) := |x|^\alpha$, $\alpha > 0$, satisfies these hypotheses. This paper proves the Markov-type inequality

$$\|P'W\|_{L_\infty(\mathbb{R})} \leqq \left\{ \int_1^{C_3 n} ds/Q^{[-1]}(s) \right\} \|PW\|_{L_\infty(\mathbb{R})},$$

degree $(P) \leqq n$. Here C_3 is some constant and $Q^{[-1]}$ is the inverse function of Q. Further, we prove the Bernstein-type inequality

$$|(PW)'(x)| \leqq C_4 \|PW\|_{L_\infty(\mathbb{R})} \left\{ \int_{Q(\max\{1,|x|\})}^{C_3 n} ds/Q^{[-1]}(s) \right\}, \qquad |x| \leqq \eta a_n,$$

and

$$|(PW)'(x)| \leqq C_5 \|PW\|_{L_\infty(\mathbb{R})} (n/a_n) \max\left\{ n^{-2/3}, 1 - \frac{|x|}{a_n} \right\}^{1/2}, \qquad |x| \geqq \eta a_n.$$

Here degree $(P) \leqq n$, η is any number in $(0, 1)$, a_n is the Mhaskar-Rahmanov-Saff number for W, and if $Q'(0)$ does not exist, x must be excluded near zero.

Key words. Freud weights, exponential weights, Markov-Bernstein inequalities

AMS(MOS) subject classifications. primary 41A17, 42C05; secondary 41A10

1. Statement of results.

Throughout $\mathcal{P}_n$ denotes the class of real polynomials of degree at most n, and $\|\cdot\|_S$ denotes the L_∞ norm over any measurable $S \subset \mathbb{R}$. Further, $C, C_1, C_2, \cdots$, denote positive constants independent of n, $P \in \mathcal{P}_n$, and $x \in \mathbb{R}$, which are not necessarily the same from line to line. We use the usual o, O notation, and $\sim$ in the following sense: If $\{c_n\}_{n=1}^\infty$ and $\{d_n\}_{n=1}^\infty$ are real sequences, then we write $c_n \sim d_n$ if there exist C_1 and C_2 such that $C_1 \leqq c_n/d_n \leqq C_2$, for n large enough. Similar notation is used for functions and sequences of functions.

The classical inequalities of Markov and Bernstein are respectively

$$\|P'\|_{[-1,1]} \leqq n^2 \|P\|_{[-1,1]}, \qquad P \in \mathcal{P}_n, \tag{1.1}$$

and

$$|P'(x)| \leqq n(1-x^2)^{-1/2} \|P\|_{[-1,1]}, \qquad P \in \mathcal{P}_n, \quad |x| < 1. \tag{1.2}$$

The interest in these inequalities (as in their weighted analogues below) lies in their application to rates of approximation by polynomials or weighted polynomials [1]. The latter in turn have wide ranging applications to establishing rates of convergence for procedures such as quadrature, interpolation, and collocation in many contexts.

* Received by the editors July 6, 1987; accepted for publication (in revised form) August 24, 1989.

† Department of Mathematics, Everyman's University, P.O. Box 39328, Tel Aviv 61392, Israel.

‡ Department of Mathematics, University of the Witwatersrand, P.O. Wits 2050, Republic of South Africa. The research of this author was completed while this author was at the National Research Institute for Mathematical Sciences, Council for Scientific and Industrial Research, P.O. Box 395, Pretoria, Republic of South Africa 0001.

Recently there has been much activity concerning weighted Markov-Bernstein inequalities for weights on $\mathbb{R}$ [3]-[8], [13]. For example, if

$$W_\alpha(x) := \exp(-|x|^\alpha), \qquad x \in \mathbb{R}, \quad \alpha > 0, \tag{1.3}$$

then for $P \in \mathcal{P}_n$,

$$\|P'W_\alpha\|_{\mathbb{R}} \leqq C\|PW_\alpha\|_{\mathbb{R}} \begin{cases} n^{1-1/\alpha} & \alpha > 1, \\ \log(n+1), & \alpha = 1, \\ 1, & 0 < \alpha < 1. \end{cases} \tag{1.4}$$

If $\alpha > 1$, we may replace $P'W_\alpha$ by $(PW_\alpha)'$ in (1.4), and if $\alpha \leqq 1$, a similar replacement is possible if we modify W_α to be differentiable at 0.

Note that (1.4) is an analogue of Markov's (1.1). In this paper, we will prove the following Markov-type generalisation of (1.4). Throughout $Q^{[-1]}(x)$ denotes the inverse function of $Q(x)$ satisfying $Q^{[-1]}(Q(x)) = x$, $x \in (0, \infty)$.

THEOREM 1.1. *Let $W(x) := e^{-Q(x)}$, where $Q(x)$ is even, continuous in $\mathbb{R}$, $Q(0) = 0$, $Q''(x)$ is continuous in $(0, \infty)$, $Q'(x)$ is positive in $(0, \infty)$ and for some $A, B > 0$,*

$$A \leqq (xQ'(x))'/Q'(x) \leqq B, \qquad x \in (0, \infty). \tag{1.5}$$

Then there exists $C > 0$ such that for $n = 1, 2, 3, \cdots$, and $P \in \mathcal{P}_n$,

$$\|P'W\|_{\mathbb{R}} \leqq \left\{ \int_1^{Cn} ds/Q^{[-1]}(s) \right\} \|PW\|_{\mathbb{R}}. \tag{1.6}$$

When $Q(x)$ satisfies a condition such as (1.5) and another condition that implies that it grows at least as fast as Cx^2, Freud [4] proved an equivalent form of (1.6). For $Q(x)$ satisfying various conditions and growing like $|x|^\alpha$, $1 < \alpha < 2$, Levin and Lubinsky [5], [6] proved (1.6). Finally, for $Q(x) := |x|$ or $Q(x)$ concave and satisfying a mild integral condition, Nevai and Totik [13] proved (1.6). However, the methods of all these authors were different, and restricted the growth scale of $Q(x)$ to lie in a fixed range of $|x|^\alpha$.

The above result simultaneously treats all possible polynomial rates of growth of $Q(x)$, and the method of proof is the same for all scales. The restriction $Q(0) = 0$ is imposed simply for convenience and can be achieved in general by replacing $W(x)$ by $W(x)/W(0)$. As examples of $Q(x)$ satisfying the above hypotheses, we mention

$$Q(x) := |x|^\alpha (\log\{A + x^2\})^\beta, \tag{1.7}$$

$\alpha > 0$, $\beta \in \mathbb{R}$, A large enough, and

$$Q(x) := |x|^{\alpha\{2 + \sin(\varepsilon \log\log(4+x^2))\}}, \tag{1.8}$$

$\alpha > 0$, ε small enough. Note that this last $Q(x)$ varies between $|x|^\alpha$ and $|x|^{3\alpha}$, and so if $\alpha = \frac{3}{4}$, for example, none of the results of [4]-[6], [13] apply to it. Similarly, for $\alpha = 1$ and $\beta \geqq -1$, the results of [4]-[6], [13] cannot treat Q of (1.7). In summary, all the L_∞ Markov-Bernstein inequalities of [4]-[6] are contained in Theorem 1.1, as are the most important ones from [13]. If, however, Q is concave and grows more slowly than any polynomial, for example, $Q(x) := (\log(1+x^2))^{1+\varepsilon}$, some $\varepsilon > 0$, then the above results do not apply, but those in [13] do. For Q of essentially faster than polynomial growth, see [7]. In the latter case, (1.6) is no longer valid.

From a historical perspective, it is interesting to note that if we replace $\|P'W\|_{\mathbb{R}}$ by $\|P'W\|_{[a,b]}$, where $[a, b]$ is any fixed finite interval, then we obtain an inequality proved by M. Dzrbasjan back in the 1950s [2, p. 398], even in greater generality. However, the constants there depend on a, b, and the methods evidently cannot be extended to all of $\mathbb{R}$.

Of course, for fixed $x \in (-1, 1)$, Bernstein's (1.2) is a better estimate than Markov's (1.1) as n increases. Further, this improved estimate is an essential ingredient of the converse theorems of polynomial approximation [1]. So we might ask whether there is an analogue for weights on $\mathbb{R}$ of Bernstein's inequality. The answer is yes, but the effect is opposite in the sense that the estimate improves with increasing $|x|$. The reason for the contrary nature is the different behaviour of functions that majorize weighted polynomials for weights on $[-1, 1]$ and on $\mathbb{R}$. To state the result, we need the following definition.

DEFINITION 1.2. Let $W(x) := e^{-Q(x)}$, where $Q(x)$ is even and continuous in $\mathbb{R}$, $Q'(x)$ exists in $(0, \infty)$, and $xQ'(x)$ is increasing in $(0, \infty)$ with limits zero and ∞ at zero and ∞, respectively. For $n = 1, 2, 3, \cdots$, we define the *Mhaskar-Rahmanov-Saff number* $a_n = a_n(W)$ to be the positive root of the equation

$$n = \frac{2}{\pi} \int_0^1 a_n t Q'(a_n t)(1 - t^2)^{-1/2}\, dt. \tag{1.9}$$

The number a_n was introduced by Mhaskar and Saff in [10], [11] and also appears in a dissertation of Rahmanov. It plays an important role in describing asymptotics associated with the orthogonal polynomials for the weight W^2 [9]. In particular, for $P \in \mathcal{P}_n$,

$$\|PW\|_{\mathbb{R}} = \|PW\|_{[-a_n, a_n]}, \tag{1.10}$$

and if P is not identically zero,

$$|PW|(x) < \|PW\|_{\mathbb{R}}, \qquad |x| > a_n. \tag{1.11}$$

In certain senses [9], [10], the interval $[-a_n, a_n]$ is asymptotically as $n \to \infty$, the smallest interval on which (1.10)-(1.11) are true. In the special case $W = W_\alpha$ (see [10, (1.3)])

$$a_n = a_n(W_\alpha) = (n/\lambda_\alpha)^{1/\alpha}, \qquad n = 1, 2, 3, \cdots, \tag{1.12}$$

where

$$\lambda_\alpha := \Gamma(\alpha) 2^{-\alpha+2} / \Gamma(\alpha/2)^2. \tag{1.13}$$

Below is our Bernstein inequality.

THEOREM 1.3. *Let $W(x)$ be as in Theorem* 1.1, *and let $a_n = a_n(W)$ for $n = 1, 2, 3, \cdots$. Let $0 < \eta < 1$ and let $b > 0$. Then for $n \geqq C_3$, $P \in \mathcal{P}_n$ and $|x| \geqq \eta a_n$,*

$$|(PW)'(x)| \leqq C_4 \|PW\|_{\mathbb{R}} \left(\frac{n}{a_n}\right) \max\left\{n^{-2/3}, 1 - \frac{|x|}{a_n}\right\}^{1/2}, \tag{1.14}$$

and for $b \leqq |x| \leqq \eta a_n$,

$$|(PW)'(x)| \leqq C_4 \|PW\|_{\mathbb{R}} \int_{Q(\max\{1, |x|\})}^{C_5 n} ds / Q^{[-1]}(s). \tag{1.15}$$

If Q' is continuous at zero, this last inequality also holds for $|x| \leqq b$.

Thus the estimate improves with increasing $|x|$, and in particular for x near the end points of the critical interval $[-a_n, a_n]$,

$$|(PW)'(x)| \leqq C_5 \|PW\|_{\mathbb{R}} n^{2/3} / a_n, \qquad P \in \mathcal{P}_n.$$

It is essential here that we consider $(PW)'$ rather than $P'W$. In fact by choosing $P \in \mathcal{P}_n$ to be a certain L_∞ extremal polynomial for W, we can show that at the largest extremum ξ_n of PW,

$$|P'W|(\xi_n) \geqq C_6 (n/a_n) \|PW\|_{\mathbb{R}},$$

where

$$\xi_n = a_n(1+0(\log n/n)^{2/3}).$$

So $|P'W|$ may be substantially larger than $|(PW)'|$ near $\pm a_n$.

If $|x| \in [\eta a_n, (1-\eta)a_n]$ for some fixed $0<\eta<\frac{1}{2}$, then both (1.14) and (1.15) yield

$$|(PW)'(x)| \leqq C_7(n/a_n)\|PW\|_{\mathbb{R}},$$

so they agree in this overlap region. For $|x|>a_n$, (1.14) may be substantially improved, but we omit this, as that region is not so important in applications. We believe that for $|x| \leqq a_n$, the estimates (1.14) and (1.15) are the "best possible" with respect to the rate of growth or decrease in n.

Let us illustrate Theorem 1.3 for $W_\alpha(x) = \exp(-|x|^\alpha)$: Let $0<\eta<1$ and $b>0$. Taking account of (1.12) and (1.13), we have for $n \geqq C_3$, $P \in \mathcal{P}_n$, and $|x| \geqq \eta(n/\lambda_\alpha)^{1/\alpha}$,

$$|(PW_\alpha)'(x)| \leqq C_7\|PW_\alpha\|_{\mathbb{R}} n^{1-1/\alpha} \max\{n^{-2/3}, 1-|x|(n/\lambda_\alpha)^{-1/\alpha}\}^{1/2}$$

and for $b \leqq |x| \leqq \eta(n/\lambda_\alpha)^{1/\alpha}$,

$$|(PW_\alpha)'(x)| \leqq C_7\|PW_\alpha\|_{\mathbb{R}} \begin{cases} n^{1-1/\alpha}, & \alpha>1, \\ \log(C_5 n/(1+|x|)), & \alpha=1, \\ (1+|x|)^{\alpha-1}, & \alpha<1. \end{cases}$$

If $\alpha>1$, this last inequality holds also for $|x| \leqq b$.

The proof of Theorems 1.1 and 1.3 involves Cauchy's integral formula for derivatives, and majorization of weighted polynomials, as well as replacement of the weight by a local approximation that is an entire function. We wish to thank E. B. Saff for pointing out that, at least in spirit, the first two ideas go back to Sewell [14]. While the methods of [5], [6], [13] worked in the L_p spaces for all $p>0$, the approach of this paper is confined to L_∞, although ideas from [8] can possibly allow passage to L_p (and even Orlicz space) inequalities.

This paper is organized as follows: In § 2, we present the main ideas in a simple case, and then present the two basic ingredients of the proof. Section 3 contains two technical lemmas, and in § 4, we apply the latter to estimate certain measures and majorisation functions. Finally, in § 5, we prove Theorems 1.1 and 1.3.

2. The two basic ingredients. The two basic ingredients of the proofs are *Cauchy's integral formula* for derivatives, and *majorisation inequalities* that bound a polynomial in the plane in terms of its norm on $\mathbb{R}$. Let us illustrate these to prove something similar to Bernstein's (1.2) for $x=0$. Let $P \in \mathcal{P}_n$ and $\varepsilon>0$. Then

$$P'(0) = \frac{1}{2\pi i}\int_{|t|=\varepsilon} \frac{P(t)}{t^2}\,dt,$$

which yields

$$|P'(0)| \leqq \varepsilon^{-1} \max_{|t|=\varepsilon} |P(t)|. \tag{2.1}$$

Next, a classical majorisation inequality of Bernstein [15] asserts that

$$|P(z)| \leqq |z+(z^2-1)^{1/2}|^n \|P\|_{[-1,1]}, \qquad z \in \mathbb{C}\backslash[-1,1]. \tag{2.2}$$

Here the branch of the square root is the principal one. Choosing $\varepsilon := n^{-1}$ in (2.1),

and applying (2.2) yields

$$\begin{aligned}|P'(0)| &\leqq n \max_{|t|=1/n} |t+(t^2-1)^{1/2}|^n \|P\|_{[-1,1]} \\ &= n(n^{-1}+(1+n^{-2})^{1/2})^n \|P\|_{[-1,1]} \\ &\leqq n e^2 \|P\|_{[-1,1]}.\end{aligned}$$

In the presence of a weight $W:\mathbb{R}\to\mathbb{R}$ that is not necessarily analytic, the replacement for (2.1) is provided by Lemma 2.1 below, and the relevant majorisation inequality is Lemma 2.2 below. The choice of ε becomes much harder, as we shall see in § 5.

LEMMA 2.1. *Let W satisfy the hypotheses of Theorem* 1.1. *Let $x\in(0,\infty)$, $\varepsilon>0$, and $P\in\mathcal{P}_n$, some $n\geqq 1$. Then*

$$|(PW)'(x)| \leqq \varepsilon^{-1} e^{\tau} \max_{|t-x|=\varepsilon} |P(t)W(|t|)|, \tag{2.3}$$

where for some C (independent of n, x, P, and ε),

$$\tau := \begin{cases} C(2x)Q'(2x)(\varepsilon/(x-\varepsilon))^2, & x\geqq 2\varepsilon, \\ 2Q(3\varepsilon)+\varepsilon Q'(x), & x<2\varepsilon. \end{cases} \tag{2.4}$$

If Q' is continuous at zero, then (2.3) *is also true for $x=0$.*

Proof. Note that x remains fixed throughout the proof. Define an entire function $\hat{W}(t) := e^{-\hat{Q}(t)}$, $t\in\mathbb{C}$, where

$$\hat{Q}(t) := Q(x)+Q'(x)(t-x), \qquad t\in\mathbb{C}.$$

Then

$$\hat{W}^{(j)}(x) = W^{(j)}(x), \qquad j=0,1,$$

so

$$\begin{aligned}|(PW)'(x)| = |(P\hat{W})'(x)| &= \left|\frac{1}{2\pi i}\int_{|t-x|=\varepsilon} \frac{(P\hat{W})(t)}{(t-x)^2}\,dt\right| \\ &\leqq \varepsilon^{-1} \max_{|t-x|=\varepsilon} |P(t)W(|t|)| \max_{|t-x|=\varepsilon} |\hat{W}(t)/W(|t|)|.\end{aligned} \tag{2.5}$$

Here,

$$|\hat{W}(t)/W(|t|)| = \exp\left(Q(|t|)-Q(x)-Q'(x)[\operatorname{Re} t-x]\right).$$

If $x<2\varepsilon$, then $|t|\leqq 3\varepsilon$, and we can use the fact that Q is strictly increasing in $(0,\infty)$ to obtain

$$|\hat{W}(t)/W(|t|)| \leqq \exp\left(2Q(3\varepsilon)+Q'(x)\varepsilon\right).$$

For this case, substituting into (2.5) yields (2.4). If $x\geqq 2\varepsilon$, then $\operatorname{Re} t\geqq x-\varepsilon\geqq\varepsilon$, and for some ξ between $\operatorname{Re} t$ and $|t|$, and η between $\operatorname{Re} t$ and x, we can write

$$\begin{aligned}|\hat{W}(t)/W(|t|)| &= \exp\left(Q(|t|)-Q(\operatorname{Re} t)+Q(\operatorname{Re} t)-Q(x)-Q'(x)(\operatorname{Re} t-x)\right) \\ &= \exp\left(Q'(\xi)(|t|-\operatorname{Re} t)+\tfrac{1}{2}Q''(\eta)(\operatorname{Re} t-x)^2\right) \\ &\leqq \exp\left(Q'(\xi)(\operatorname{Im} t)^2/\operatorname{Re} t+\tfrac{1}{2}|Q''(\eta)|\varepsilon^2\right) \\ &\qquad\qquad \text{(by the inequality } (a^2+b^2)^{1/2}-a\leqq b^2/a,\ a>0,\ b\geqq 0) \\ &\leqq \exp\left(\xi Q'(\xi)(\varepsilon/(x-\varepsilon))^2+\eta^2|Q''(\eta)|(\varepsilon/(x-\varepsilon))^2\right),\end{aligned} \tag{2.6}$$

since Re $t, \xi, \eta \geqq x-\varepsilon$, while $|\text{Im } t| \leqq \varepsilon$. Next, we will need two immediate consequences of (1.5):

(2.7) $sQ'(s)$ is positive and strictly increasing in $(0, \infty)$;

(2.8) $$(A-1)Q'(s)/s \leqq Q''(s) \leqq (B-1)Q'(s)/s, \qquad s \in (0, \infty).$$

The first yields

$$\xi Q'(\xi) \leqq (x+\varepsilon)Q'(x+\varepsilon) \leqq 2xQ'(2x),$$

and the first and second yield for some C_3,

$$|\eta^2 Q''(\eta)| \leqq C_3(x+\varepsilon)Q'(x+\varepsilon) \leqq C_3 2xQ'(2x).$$

Substituting into the right-hand side of (2.6) and then into (2.5) yields (2.3). □

To bound $\max_{|t-x|=\varepsilon} |P(t)W(|t|)|$, we will use a majorisation inequality of the form

(2.9) $$|P(t)W(|t|)| \leqq \|PW\|_{\mathbb{R}} \exp(nU_n(t/a_n)).$$

The precise technical details are contained in the following lemma.

LEMMA 2.2. *Let $W := e^{-Q}$ satisfy the hypotheses of Theorem* 1.1. *For $n \geqq 1$, let $a_n = a_n(W)$ be defined by* (1.9), *and for almost every $x \in (-1, 1)$, let*

(2.10) $$\mu_n(x) := 2\pi^{-2} \int_0^1 \frac{(1-x^2)^{1/2}}{(1-s^2)^{1/2}} \frac{a_n sQ'(a_n s) - a_n xQ'(a_n x)}{n(s^2-x^2)}\, ds.$$

Further, let

(2.11) $$\chi_n := 2\pi^{-1} \int_0^1 \frac{Q(a_n t)}{(1-t^2)^{1/2}}\, dt + n \log 2,$$

and for $z \in \mathbb{C}$, let

(2.12) $$U_n(z) := \int_{-1}^1 \log|z-t| \mu_n(t)\, dt - Q(a_n|z|)/n + \chi_n/n.$$

Then

(a) *For almost every $x \in (-1, 1)$,*

(2.13) $$0 < \mu_n(x) < \infty,$$

and

(2.14) $$\int_{-1}^1 \mu_n(x)\, dx = 1.$$

(b) *For $n \geqq 1$,*

(2.15) $$U_n(x) = 0, \qquad x \in [-1, 1],$$

and

(2.16) $$U_n(x) < 0; \qquad U_n'(x) < 0, \qquad x \in (1, \infty).$$

Furthermore, for some C_3, C_4, ε_0,

(2.17) $$-C_3\varepsilon^{3/2} \leqq U_n(1+\varepsilon) \leqq -C_4\varepsilon^{3/2}, \ \varepsilon \in [0, \varepsilon_0],$$

and given $\delta > 0$, there exists $C_4 = C_4(\delta)$ such that

(2.18) $$U_n(x) \leqq -C_4 \log x, \ x \in [1+\delta, \infty).$$

(c) *For $n \geqq 1$, $P \in \mathcal{P}_n$ and $z \in \mathbb{C}$,*

(2.19) $$|P(z)W(|z|)| \leqq \|PW\|_{[-a_n, a_n]} \exp(nU_n(z/a_n)).$$

Furthermore

$$\|PW\|_{\mathbb{R}} = \|PW\|_{[-a_n, a_n]}, \tag{2.20}$$

and if P is not identically zero,

$$|(PW)(x)| < \|PW\|_{\mathbb{R}}, \qquad |x| > a_n. \tag{2.21}$$

Proof. (a) These follow from (5.40) and (5.41) in [9, pp. 37-38] with $R = a_n$, $\mu_n = \mu_{n,a_n}$ in the notation of [9].

(b) First, (2.15) is a restatement of (5.45) in [9, p. 38] and (2.16) is a restatement of (5.56) in [9, p. 39]. Next, (2.17) is the special case $R = a_n$ of (6.29) in [9, p. 45]. Finally for (2.18), we refer the reader to [9, p. 55, line 11] where it is shown that

$$\begin{aligned} U_n(x) &\leqq -[C_5 + C_6 \log(x/(1+\delta))], \qquad x \geqq 1+\delta, \\ &\leqq -C_7 \log x, \qquad x \geqq 1+\delta. \end{aligned}$$

(c) This is Theorem 7.1(i), (ii) in [9, pp. 49-50] with $R = a_n$. □

3. Two technical lemmas. Below are some technical estimates involving Q.

LEMMA 3.1. *Let $W := e^{-Q}$ satisfy the hypotheses of Theorem* 1.1. *Then*

(a)

$$Q'(1)x^{B-1} \leqq Q'(x) \leqq Q'(1)x^{A-1}, \qquad x \in (0, 1], \tag{3.1}$$

and

$$Q'(1)x^{A-1} \leqq Q'(x) \leqq Q'(1)x^{B-1}, \qquad x \in [1, \infty). \tag{3.2}$$

(b)

$$t^A \leqq tQ'(tx)/Q'(x) \leqq t^B, \qquad x \in (0, \infty), \qquad t \in (1, \infty), \tag{3.3}$$

and

$$AQ(x) \leqq xQ'(x) \leqq BQ(x), \qquad x \in (0, \infty). \tag{3.4}$$

(c) *For $n = 1, 2, 3, \cdots$, there is a unique positive root a_n of* (1.9), *and for some $C_3, C_4 > 0$,*

$$n^{C_3} \leqq a_n \leqq n^{C_4}, \ n \text{ large enough.} \tag{3.5}$$

Finally, if $0 < a < b < \infty$, then uniformly for $x \in [a, b]$ and n large enough,

$$a_n x Q'(a_n x) \sim Q(a_n x) \sim n. \tag{3.6}$$

Proof. (a) Let $r > s > 0$. Then

$$\frac{rQ'(r)}{sQ'(s)} = \exp\left(\int_s^r \frac{(uQ'(u))'}{uQ'(u)}\, du\right),$$

and so (1.5) shows that

$$\exp\left(A\int_s^r \frac{du}{u}\right) \leqq \frac{rQ'(r)}{sQ'(s)} \leqq \exp\left(B\int_s^r \frac{du}{u}\right),$$

that is

$$(r/s)^A \leqq \frac{rQ'(r)}{sQ'(s)} \leqq (r/s)^B. \tag{3.7}$$

Choosing $s := 1$ and $r := x > 1$ yields (3.2). Choosing $s := x \in (0, 1]$ and $r := 1$ yields (3.1).

(b) Choosing $r := tx$ and $s := x$ with $t \in (1, \infty)$ in (3.7) yields (3.3). Integrating (1.5) from zero to x yields (3.4), since $Q(0) = 0$ and (3.1) ensures that

$$\lim_{x \to 0+} xQ'(x) = 0. \tag{3.8}$$

(c) The monotonicity of $tQ'(t)$ (see (2.7)) yields

$$n \leqq \left(\frac{2}{\pi} \int_0^1 \frac{dt}{\sqrt{1-t^2}} \right) a_n Q'(a_n) = a_n Q'(a_n),$$

and

$$n \geqq \left(\frac{2}{\pi} \int_{1/2}^1 \frac{dt}{\sqrt{1-t^2}} \right) \frac{a_n}{2} Q'\left(\frac{a_n}{2} \right) = \frac{a_n}{3} Q'\left(\frac{a_n}{2} \right).$$

It then follows from (3.3) that

$$a_n Q'(a_n) \sim n, \qquad n \geqq 1. \tag{3.9}$$

In view of (3.2), we obtain (3.5). Finally, from (3.3), (3.4), and (3.9), we deduce (3.6). □

In estimating μ_n, we will also need the following lemma.

LEMMA 3.2. *Let $W := e^{-Q}$ satisfy the hypotheses of Theorem* 1.1. *Let $0 < r < 2$. Define*

$$\psi_n(x) := \int_x^2 \frac{a_n Q'(a_n t)}{nt}\, dt, \qquad x \in (0, 2), \qquad n \geqq 1. \tag{3.10}$$

(a) *Then $\psi_n(x)$ is a positive decreasing function in $(0, 2)$, satisfying*

$$\lim_{x \to 0+} x\psi_n(x) = 0. \tag{3.11}$$

(b) *There exists C such that*

$$\psi_n(x) \geqq C[1 + a_n Q'(a_n x)/n], \qquad x \in (0, r], \tag{3.12}$$

and

$$\psi_n(x) \geqq C[1 + Q(a_n x)/(nx)], \qquad x \in (0, r]. \tag{3.13}$$

(c) *Furthermore, uniformly for $x \in (0, r/2]$ and $n \geqq 1$,*

$$\psi_n(x) \sim \psi_n(2x). \tag{3.14}$$

(d) *Given $\delta > 0$, there exists C_1 such that*

$$x\psi_n(x) \geqq C_1/n, \qquad x \in [\delta/a_n, r]. \tag{3.15}$$

For $n \geqq 1$ and $x \in (0, r]$,

$$\psi_n(x) = \frac{a_n}{n} \int_{Q(a_n|x|)}^{Q(2a_n)} \frac{dv}{Q^{[-1]}(v)}. \tag{3.16}$$

(e) *Let $\delta_n \in (0, 1]$ be the smallest number such that*

$$t\psi_n(t) \geqq 1/n, \qquad t \in [\delta_n, 1]. \tag{3.17}$$

Then δ_n exists for n large enough,

$$\delta_n \psi_n(\delta_n) = 1/n, \tag{3.18}$$

and for some C_2,

$$\delta_n \leqq C_2/a_n. \tag{3.19}$$

If in addition,

$$\int_0^1 Q'(s)/s\,ds < \infty, \tag{3.20}$$

then given $\beta > 0$, *we have for n large enough,*

$$\psi_n(x) \sim \psi_n(1/a_n) \sim \psi_n(\delta_n), \text{ uniformly for } x \in (0, \beta/a_n]. \tag{3.21}$$

Proof. (a) Given $0 < \eta < 2$, we have for $x \in (0, \eta)$,

$$\begin{aligned}\psi_n(x) &\leqq \int_x^\eta \frac{a_n t Q'(a_n t)}{nt^2}\,dt + \int_\eta^2 \frac{a_n Q'(a_n t)}{nt}\,dt \\ &\leqq a_n \eta Q'(a_n\eta)(nx)^{-1} + \int_\eta^2 \frac{a_n Q'(a_n t)}{nt}\,dt,\end{aligned}$$

by the monotonicity of $uQ'(u)$. We deduce that

$$0 \leqq \limsup_{x\to 0+} x\psi_n(x) \leqq a_n\eta Q'(a_n\eta)/n.$$

Letting $\eta \to 0+$ and using (3.8) yields (3.11).

(b) Let $0 < r < s < 2$. Now for $x \in (0, r]$,

$$\begin{aligned}\psi_n(x) &\geqq a_n x Q'(a_n x)\int_x^s \frac{dt}{nt^2} + \int_s^2 \frac{a_n Q'(a_n t)}{nt}\,dt \\ &\geqq C_3[a_n Q'(a_n x)/n + 1], \qquad x \in (0, r],\end{aligned}$$

by (3.6). This yields (3.12) and then (3.13) follows from (3.4).

(c) To prove (3.14), we note that

$$\begin{aligned}\psi_n(2x) &\leqq \psi_n(x) = \psi_n(2x) + \int_x^{2x} \frac{a_n Q'(a_n t)}{nt}\,dt \\ &\leqq \psi_n(2x) + 2a_n x Q'(2a_n x)\int_x^{2x} \frac{dt}{nt^2} \\ &= \psi_n(2x) + a_n Q'(2a_n x)/n \\ &\leqq \psi_n(2x)(1 + 1/C),\end{aligned}$$

by (3.12).

(d) From (3.12) and the monotonicity of $uQ'(u)$, we can easily deduce (3.15) for $x \in [\delta/a_n, r]$. Next, the substitution $v = Q(a_n t)$ in (3.10) yields (3.16).

(e) First, the existence of δ_n for n large enough follows from the continuity and positivity of ψ_n in $(0, 1]$, and from (3.11) and (3.12). Next, (3.18) is immediate, while (3.13) ensures that

$$1/n = \delta_n \psi_n(\delta_n) \geqq CQ(a_n\delta_n)/n,$$

which in turn implies (3.19) as $\lim_{x\to\infty} Q(x) = \infty$.

Finally, to prove (3.21) assuming (3.20), we note that for $x \in (0, \beta/a_n]$,

$$\psi_n(\beta/a_n) \leqq \psi_n(x) = \int_x^{\beta/a_n} \frac{a_n Q'(a_n t)}{nt}\,dt + \psi_n(\beta/a_n)$$

$$= \frac{a_n}{n} \int_{a_n x}^{\beta} \frac{Q'(s)}{s}\, ds + \psi_n(\beta/a_n)$$

$$\leqq C_7 \left[Q\left(a_n \cdot \frac{1}{a_n}\right) \Big/ \left(n \cdot \frac{1}{a_n}\right) + \psi_n\left(\frac{1}{a_n}\right) \right] \qquad \text{(by (3.20) and (3.14))}$$

$$\leqq C_8 \left[\psi_n\left(\frac{1}{a_n}\right) + \psi_n\left(\frac{1}{a_n}\right) \right],$$

by (3.13). Applying (3.14) again yields the first $\sim$ in (3.21). The second follows from (3.19). □

4. Estimates for μ_n and U_n. To estimate U_n, we first need two simple estimates for μ_n, as follows.

LEMMA 4.1. *Let $W := e^{-Q}$ satisfy the hypotheses of Theorem* 1.1, *and let ψ_n be defined by* (3.10) *for $n \geqq 1$. Let $0 < \eta < 1$. Then there exists C such that for n large enough,*

$$\mu_n(x) \leqq C(1-x^2)^{1/2}, \qquad x \in [\eta, 1], \tag{4.1}$$

and

$$\mu_n(x) \leqq C\psi_n(x), \qquad x \in (0, \eta]. \tag{4.2}$$

Proof. Recall from (2.10) that

$$\mu_n(x) = 2\pi^{-2} \int_0^1 \frac{(1-x^2)^{1/2}}{(1-s^2)^{1/2}} \frac{\Delta}{s+x} \frac{a_n}{n}\, ds, \tag{4.3}$$

where

$$\begin{aligned} \Delta := \Delta(n, s, x) &:= \frac{a_n s Q'(a_n s) - a_n x Q'(a_n x)}{a_n s - a_n x} \\ &= (uQ'(u))', \end{aligned} \tag{4.4}$$

for some u between $a_n s$ and $a_n x$. Then by (1.5),

$$\Delta \leqq BQ'(u) \leqq Ba_n r Q'(a_n r)/(a_n t), \tag{4.5}$$

where $r := \max\{s, x\}$; $t := \min\{s, x\}$, and we have used the monotonicity of $uQ'(u)$.

Suppose first for some $0 < \eta < 1$, $x \in [\eta, 1]$. Then (4.5) and (3.6) show that uniformly for such x, and for $s \in [\eta/2, 1]$, and for n large enough

$$\Delta \leqq Cn/a_n.$$

Furthermore, for $s \in [0, \eta/2]$, the positivity and monotonicity of $uQ'(u)$ yield

$$\Delta \leqq a_n x Q'(a_n x)/(a_n x - a_n \eta/2) \leqq Cn/a_n,$$

by (3.6) again. These estimates and (4.3) and the fact that $s + x \geqq \eta$ yield (4.1) for $x \in [\eta, 1]$.

Next, suppose $x \in (0, \frac{1}{4}]$. Then for $s \in [0, x/2]$,

$$\Delta \leqq a_n x Q'(a_n x)/(a_n x - a_n x/2) = 2Q'(a_n x),$$

and for $s \in [x/2, 2x]$, by (4.5),

$$\Delta \leqq Ba_n 2x Q'(a_n 2x)/(a_n x/2) = 4BQ'(2a_n x).$$

Finally, for $s \in [2x, 1]$, as $s - x \geqq s/2$,

$$\Delta \leqq a_n s Q'(a_n s)/(a_n s/2) = 2Q'(a_n s).$$

Hence

$$\mu_n(x) \leqq 4\pi^{-2} Q'(a_n x) \int_0^{x/2} \frac{(1-x^2)^{1/2}}{(1-s^2)^{1/2}} \frac{1}{x} \frac{a_n}{n}\, ds$$
$$+ 8\pi^{-2} BQ'(2a_n x) \int_{x/2}^{2x} \frac{(1-x^2)^{1/2}}{(1-s^2)^{1/2}} \frac{1}{x} \frac{a_n}{n}\, ds + 4\pi^{-2} \int_{2x}^{1} \frac{(1-x^2)^{1/2}}{(1-s^2)^{1/2}} \frac{Q'(a_n s)}{s} \frac{a_n}{n}\, ds$$
$$\leqq C_1 \left[Q'(2a_n x) a_n / n + \int_{2x}^{1/2} \frac{Q'(a_n s)}{ns}\, a_n\, ds + \int_{1/2}^{1} \frac{1}{(1-s^2)^{1/2}} \frac{n}{s^2 n}\, ds \right].$$

Here we have used the monotonicity of $uQ'(u)$, as well as the fact that $(1-s^2)^{1/2} \geqq C$, $s \in [0, \frac{1}{2}]$, and we have also used (3.6). Thus

$$\mu_n(x) \leqq C_2[Q'(2a_n x) a_n / n + \psi_n(2x) + 1] \leqq C_3 \psi_n(2x),$$

by (3.12). Then (3.14) yields (4.2) for $x \in (0, \frac{1}{4}]$. If $\frac{1}{4} \leqq \eta < 1$, then (4.1) yields (4.2) in the range $[\eta', \eta]$ for any $0 < \eta' < \frac{1}{4}$. □

We can now estimate $U_n(x+iy)$ for x in a neighbourhood of zero, as follows.

LEMMA 4.2. *Let W satisfy the hypotheses of Theorem* 1.1, *and let ψ_n be defined by* (3.10) *for $n \geqq 1$. Let $0 < \eta < 1$. Then there exists C_4 such that for n large enough, $x \in [-\eta, \eta]$ and $y \in [-1, 1]$,*

$$U_n(x+iy) \leqq C_4 \psi_n(x)|y|. \tag{4.6}$$

Furthermore, if Q satisfies (3.20), *and $\{\delta_n\}_1^\infty \subset (0,1)$ satisfies* (3.17) *and* (3.18), *then given $K > 0$, we have*

$$U_n(x+iy) \leqq C_4 n^{-1}, \tag{4.7}$$

for $|x|, |y| \leqq K\delta_n$ and n large enough.

Proof. Since $U_n(x)$ is even, we may assume $x \in [0, \eta]$. We may also assume $x \neq 0$, in view of the continuity of U_n and monotonicity of ψ_n. Then by (2.12) and (2.15),

$$U_n(x+iy) = U_n(x+iy) - U_n(x)$$
$$= \left(\frac{1}{2}\right) \int_{-1}^{1} \log\left(1 + \left(\frac{y}{x-t}\right)^2\right) \mu_n(t)\, dt + \left\{ \frac{Q(a_n|x|) - Q(a_n(x^2+y^2)^{1/2})}{n} \right\} \tag{4.8}$$
$$\leqq \int_0^1 \log\left(1 + \left(\frac{y}{x-t}\right)^2\right) \mu_n(t)\, dt,$$

by monotonicity of Q, evenness of μ_n, and the fact that

$$(x-t)^2 \leqq (x+t)^2, \qquad t \in [0,1].$$

Now choose $\eta' \in (\eta, 1)$. By the estimates of Lemma 4.1,

$$U_n(x+iy) \leqq \left(\int_0^{x/2} + \int_{x/2}^{\eta'} + \int_{\eta'}^{1} \right) \log\left(1 + \left(\frac{y}{x-t}\right)^2\right) \mu_n(t)\, dt$$
$$\leqq C_3 \log\left(1 + \left(\frac{2y}{x}\right)^2\right) \int_0^{x/2} \psi_n(t)\, dt$$
$$+ C_3 \psi_n(x/2) \int_{x/2}^{\eta'} \log\left(1 + (y/(x-t))^2\right) dt$$
$$+ C_3 \log\left(1 + \left(\frac{y}{\eta' - \eta}\right)^2\right) \int_{\eta'}^{1} (1-t^2)^{1/2}\, dt,$$

where we have also used the monotonicity of ψ_n. Making the substitution $x-t=uy$ in the second integral on the last right-hand side, and using the inequality

$$\log(1+u) \leqq u, \qquad u \in (0,\infty), \tag{4.9}$$

in the third term on the last right-hand side, we obtain for $x \in (0, \eta]$, $y \in \mathbb{R}$, and n large enough,

$$U_n(x+iy) \leqq C_3 \log(1+(2y/x)^2) \int_0^{x/2} \psi_n(t)\, dt + C_5\{\psi_n(x/2)|y|+y^2\}. \tag{4.10}$$

Here, by the definition (3.10) of $\psi_n(t)$, and then by interchanging integrals,

$$\begin{aligned}\int_0^{x/2} \psi_n(t)\, dt &= \int_0^{x/2} \int_t^2 \frac{a_n Q'(a_n s)}{ns}\, ds\, dt \\ &= \int_0^{x/2} \left(\int_0^s dt\right) \frac{a_n Q'(a_n s)}{ns}\, ds + \int_{x/2}^2 \left(\int_0^{x/2} dt\right) \frac{a_n Q'(a_n s)}{ns}\, ds \\ &= Q(a_n x/2)/n + (x/2)\psi_n(x/2) \\ &\leqq C_6(x/2)\psi_n(x/2),\end{aligned}$$

by (3.13). Note too that $u \log(1+u^{-2})$ is bounded in $(0,\infty)$, and hence

$$\log(1+(2y/x)^2) \leqq C_7|y|/x.$$

Substituting into (4.10), and then using (3.13) and (3.14), we obtain (4.6) for $x \in (0, \eta]$ and $|y| \leqq 1$. Since U_n is even, (4.6) follows for $x \in [-\eta, \eta]$ and $|y| \leqq 1$.

Next, if Q satisfies (3.20), then for $|x|$, $|y| \leqq \delta_n$, which implies $|x|$, $|y| \leqq C/a_n$ (by (3.19)), then by (4.6),

$$\begin{aligned}U_n(x+iy) &\leqq C_8|y|\psi_n(x) \\ &\leqq C_9\delta_n\psi_n(1/a_n) \qquad \text{(by (3.14) and (3.21))} \\ &\leqq C_{10}\delta_n\psi_n(\delta_n) = C_{10}/n,\end{aligned}$$

as ψ_n is decreasing, and by (3.14) and (3.19). □

Next, we consider the range $|x| \in [\eta, 1]$, as follows.

LEMMA 4.3. *Let $W := e^{-Q}$ satisfy the hypotheses of Theorem* 1.1. *Let $0<\eta<1$. Then there exists C_4 such that for $|x| \in [\eta, 1]$, $|y| \leqq 1$, and n large enough,*

$$U_n(x+iy) \leqq C_4 \max\{|y|^{3/2}, |y|(1-|x|)^{1/2}\}. \tag{4.11}$$

Proof. First, from (4.1),

$$\mu_n(t) \leqq C_5(1-t)^{1/2}, \qquad t \in [\eta/2, 1],$$

and so for $x \in [\eta, 1]$, (4.8) yields,

$$\begin{aligned}U_n(x+iy) &\leqq \left(\int_0^{\eta/2} + \int_{\eta/2}^1\right) \log\left(1+\left(\frac{y}{x-t}\right)^2\right) \mu_n(t)\, dt \\ &\leqq \log\left(1+\left(\frac{2y}{\eta}\right)^2\right) \int_0^{\eta/2} \mu_n(t)\, dt\end{aligned}$$

$$+C_5\int_{\eta/2}^{1}\log\left(1+(y/(x-t))^2\right)(1-t)^{1/2}\,dt \qquad \text{(by (4.1))}$$

$$\leqq\left(\frac{2y}{\eta}\right)^2+C_5|y|\int_{(x-1)/|y|}^{(x-\eta/2)/|y|}\log\left(1+u^{-2}\right)(1-(x-u|y|))^{1/2}\,du,$$

by (2.14), (4.9), and the substitution $x-t=u|y|$. Here

$$(1-(x-u|y|))^{1/2}\leqq(1-x)^{1/2}+(|uy|)^{1/2},$$

so

$$U_n(x+iy)\leqq(2y/\eta)^2+C_5|y|(1-x)^{1/2}\int_{-\infty}^{\infty}\log\left(1+u^{-2}\right)du$$

$$+C_5|y|^{3/2}\int_{-\infty}^{\infty}\log\left(1+u^{-2}\right)|u|^{1/2}\,du.$$

Then (4.11) follows. □

Finally, we turn to the range $|x|\in[1,\infty)$, as follows.

LEMMA 4.4. *Let W satisfy the hypotheses of Theorem* 1.1. *Then for $|x|\in[1,2]$, $|y|\leqq 1$, and n large enough,*

$$U_n(x+iy)\leqq C_4\{|y|^{3/2}-C_5(x-1)^{3/2}\}, \tag{4.12}$$

while for $|x|\in[2,\infty)$, $|y|\leqq 1$, and n large enough,

$$U_n(x+iy)\leqq C_4\{|y|^{3/2}-C_5\log|x|\}. \tag{4.13}$$

Proof. We may assume $x\in[1,\infty)$. Then

$$\begin{aligned}U_n(x+iy)&=U_n(x+iy)-U_n(x)+U_n(x)\\&=\left(\frac{1}{2}\right)\int_{-1}^{1}\log\left\{1+\left(\frac{y}{x-t}\right)^2\right\}\mu_n(t)\,dt\\&\qquad+\{Q(a_n|x|)-Q(a_n(x^2+y^2)^{1/2})\}/n+U_n(x)\\&\leqq\int_{0}^{1}\log\left\{1+\left(\frac{y}{1-t}\right)^2\right\}\mu_n(t)\,dt+U_n(x).\end{aligned} \tag{4.14}$$

A glance at the proof of Lemma 4.3 shows that our estimate for $U_n(1+iy)$ (that is, $x=1$ there) is an estimate for the integral in this last right-hand side. Thus the proof of Lemma 4.3 yields an upper bound of $C_5|y|^{3/2}$ for the integral. Next, (2.16), (2.17), and the monotonicity of U_n in $(1,\infty)$ show that

$$U_n(x)\leqq-C_6(x-1)^{3/2},\qquad x\in[1,2],$$

and by (2.18),

$$U_n(x)\leqq-C_6\log x,\qquad x\in[2,\infty).$$

Then (4.12) and (4.13) follow. □

5. Proof of the results of § 1. We first show that for most purposes we can replace our weight W by a very similar weight W^* that is twice continuously differentiable at zero.

LEMMA 5.1. *Let $W := e^{-Q}$ satisfy the hypotheses of Theorem* 1.1. *Let $\rho > 0$. Then there exists a weight $W^* := e^{-Q^*}$ with the following properties*:

(a) *Q^* is even and twice continuously differentiable in $\mathbb{R}$.*

(b) *$Q^{*\prime}$ is positive in $(0, \infty)$, and for some $A^*, B^* > 0$,*

$$A^* \leqq (xQ^{*\prime}(x))'/Q^{*\prime}(x) \leqq B^*, \qquad x \in (0, \infty). \tag{5.1}$$

(c)

$$Q^*(x) = Q(x), \qquad |x| \geqq \rho, \tag{5.2}$$

$$Q^{*[-1]}(y) = Q^{[-1]}(y), \qquad y \geqq Q(\rho), \tag{5.3}$$

and

$$W^*(x) \sim W(x), \qquad x \in \mathbb{R}. \tag{5.4}$$

Furthermore, (3.20) holds for Q^*.

Proof. Define

$$Q^*(x) := Q(x), \qquad |x| \geqq \rho.$$

In $[-\rho, \rho]$, we will define Q^* so that Q^* is twice continuously differentiable there and satisfies (5.1) in $(0, \infty)$. Let ε be a small positive number, let

$$L(x) := \{x^2 + \varepsilon(x^2 - \rho^2)^4\}^{1/2}, \qquad x \in [-\rho, \rho],$$

and let

$$Q^*(x) := Q(L(x)), \qquad x \in (-\rho, \rho).$$

Then $Q^*(x)$ is even and twice continuously differentiable in $(-\rho, \rho)$ since $L(x)$ is bounded below there by a positive number. As

$$L(\rho) = \rho, \qquad L'(\rho) = 1, \qquad L''(\rho) = 0,$$

we see that $Q^{*\prime\prime}(x)$ is continuous at ρ and so continuous in $\mathbb{R}$. The properties (a) and (c) above now follow. It remains to show that for small enough ε, Q^* satisfies (5.1). First,

$$L'(x) = \frac{x}{L(x)}\{1 + 4\varepsilon(x^2 - \rho^2)^3\} = \frac{x}{L(x)}\{1 + O(\varepsilon)\},$$

uniformly for $x \in [-\rho, \rho]$, as $\varepsilon \to 0+$. Furthermore, we see that

$$\frac{xL''(x)}{L'(x)} = 1 - \left(\frac{x}{L(x)}\right)^2 + O(\varepsilon),$$

uniformly for $x \in [-\rho, \rho]$, as $\varepsilon \to 0$. Then provided ε is small enough, we see that $L'(x) > 0$, so $Q^{*\prime}(x) > 0$, $x \in (0, \rho)$. Furthermore, a straightforward calculation shows that

$$\begin{aligned}(xQ^{*\prime}(x))'/Q^{*\prime}(x) &= 1 + \frac{xL'(x)}{L(x)} \frac{(uQ'(u))'}{Q'(u)}_{|u=L(x)} + \frac{xL''(x)}{L'(x)} - \frac{xL'(x)}{L(x)} \\ &= 2\left(1 - \left(\frac{x}{L(x)}\right)^2\right) + \left(\frac{x}{L(x)}\right)^2 \frac{(uQ'(u))'}{Q'(u)}_{|u=L(x)} + O(\varepsilon),\end{aligned} \tag{5.5}$$

uniformly for $x \in (0, \rho]$ as $\varepsilon \to 0+$. Since

$$x/L(x) \leqq 1, \qquad x \in [0, \rho],$$

(1.5) shows that the extreme right-hand side of (5.5) is bounded above by $2+B+1$, for $x\in[0,\rho]$, and ε small enough. Furthermore, if

$$x/L(x)\leqq \tfrac{1}{2},$$

we see from (1.5) that the extreme right-hand side of (5.5) is bounded below by $1+O(\varepsilon)$, while if

$$x/L(x)\geqq \tfrac{1}{2},$$

it is bounded below by $A/4+O(\varepsilon)$. Thus (5.1) is fulfilled for $x\in(0,\rho]$ and so in $(0,\infty)$. Finally as $Q^{*\prime}(x)/x$ is bounded near zero, (3.20) follows. □

Because of problems caused by Q not being sufficiently smooth at zero, we have to break the proof of Theorems 1.1 and 1.3 into several stages.

LEMMA 5.2. *Let W satisfy the hypotheses of Theorem* 1.1, *let $0<\eta<1$, and ψ_n and $\{\delta_n\}_1^\infty$ be as in* (3.10) *and* (3.17), *respectively. Then there exist C_7 and C_8 such that*

$$|(PW)'(x)|\leqq C_7\|PW\|_{\mathbb{R}}\psi_n(|x|/a_n)n/a_n, \tag{5.6}$$

for $P\in\mathcal{P}_n$, $n\geqq C_8$, and $|x|\in[\delta_n a_n,\eta a_n]$.

Proof. Let $s:=x/a_n\in[\delta_n,\eta]$ and $\varepsilon:=a_n/(2n\psi_n(s))$. We will estimate the quantities τ and $\max_{|t-x|=\varepsilon}|P(t)W(|t|)|$ in (2.3) of Lemma 2.1. First, by (3.17), $\varepsilon=x/(2ns\psi_n(s))\leqq x/2$. Then by (2.4),

$$\begin{aligned}\tau &= C2xQ'(2x)(\varepsilon/(x-\varepsilon))^2\\ &\leqq C2a_nsQ'(2a_ns)(2\varepsilon/x)^2\\ &\leqq C_1ns\psi_n(s)(2/[2n\psi_n(s)s])^2\leqq C_2,\end{aligned} \tag{5.7}$$

by (3.12), (3.14), the choice of ε, and (3.17). Next, by (2.19),

$$\max_{|t-x|=\varepsilon}|P(t)W(|t|)|\leqq\|PW\|_{[-a_n,a_n]}\max_{|t-x|=\varepsilon}\exp(nU_n(t/a_n)). \tag{5.8}$$

Now for $|t-x|=\varepsilon$, we have

$$\operatorname{Re} t/a_n\geqq(x-\varepsilon)/a_n\geqq x/(2a_n)=s/2,\qquad |\operatorname{Im} t|/a_n\leqq\varepsilon/a_n\leqq x/(2a_n)=s/2.$$

Then (4.6) shows that

$$U_n(t/a_n)\leqq C_3\psi_n(\operatorname{Re} t/a_n)\varepsilon/a_n\leqq C_4\psi_n(s/2)(n\psi_n(s))^{-1}\leqq C_5n^{-1},$$

by monotonicity of ψ_n, (3.14), and choice of ε. Thus

$$\max_{|t-x|=\varepsilon}\exp(nU_n(t/a_n))\leqq C_6, \tag{5.9}$$

and substituting the estimates (5.7) to (5.9) in (2.3) yields (5.6) for $x\in[a_n\delta_n,a_n\eta]$. As W is even, (5.6) also follows for $|x|\in[a_n\delta_n,a_n\eta]$. □

LEMMA 5.3. *Let W satisfy the hypotheses of Theorem* 1.1, *with the additional restriction that* (3.20) *holds. Let $0<\eta<1$, and let ψ_n and $\{\delta_n\}_1^\infty$ be as in* (3.10) *and* (3.17), *respectively. There exist C_5 and C_6 such that*

$$|(PW)'(x)|\leqq C_5\|PW\|_{\mathbb{R}}\psi_n(|x|/a_n)n/a_n, \tag{5.10}$$

for $P\in\mathcal{P}_n$, $n\geqq C_6$, and $|x|\leqq\delta_n a_n$.

Proof. Let $s:=x/a_n\in[0,\delta_n]$ and $\varepsilon:=a_n/(n\psi_n(\delta_n))=a_n\delta_n$, n large enough. Again we will estimate τ and $\max_{|t-x|=\varepsilon}\exp(nU_n(t/a_n))$. It is an easy consequence of (3.20) and (3.3) that

$$\lim_{s\to0+}Q'(s)=0. \tag{5.11}$$

Next

$$\varepsilon, x \leqq a_n\delta_n \leqq C_1,$$

by (3.19), so from (2.4),

$$\tau = 2Q(3\varepsilon) + \varepsilon Q'(x) \leqq C_2.$$

Furthermore, if $|t-x| = \varepsilon$, then $|t| \leqq 2\delta_n a_n$, so by (4.7) in Lemma 4.2,

$$\max_{|t-x|=\varepsilon} (nU_n(t/a_n)) \leqq C_3.$$

Then (2.3) and (5.8) yield (5.10) but with $\psi_n(\delta_n)$ replacing $\psi_n(|x|/a_n)$ in the right-hand side of (5.10). However, (3.19) and (3.21) yield

$$\psi_n(|x|/a_n) \sim \psi_n(\delta_n) \sim \psi_n(1/a_n), |x| \leqq \max\{\delta_n a_n, 1\}. \tag{5.12}$$

So (5.10) holds as stated. □

We can now proceed to the proof of (1.15) of Theorem 1.3.

Proof of (1.15) *of Theorem* 1.3. Let $0 < \eta < 1$. In the case where (3.20) holds, Lemmas 5.2 and 5.3 yield

$$|(PW)'(x)| \leqq C_7 \|PW\|_{\mathbb{R}} \psi_n(|x|/a_n) n/a_n, \tag{5.13}$$

for $P \in \mathcal{P}_n$, $|x| \leqq \eta a_n$, and $n \geqq C_8$. By (3.16) and (5.12),

$$\begin{aligned} \psi_n(|x|/a_n)n/a_n &\sim \psi_n(\max\{1,|x|\}/a_n)n/a_n \\ &= \int_{Q(\max\{1,|x|\})}^{Q(2a_n)} dv/Q^{[-1]}(v) \sim \int_{Q(\max\{1,|x|\})}^{Cn} dv/Q^{[-1]}(v), \end{aligned} \tag{5.14}$$

by (3.6), provided only C is large enough. (Recall that $|x| \leqq \eta a_n < a_n$.) Then (5.13) yields (1.15) for $|x| \leqq \eta a_n$, provided (3.20) holds.

Next, we turn to the case where (3.20) does not hold. Let $\rho > 0$ and let $W^* = e^{-Q^*}$ be the weight in Lemma 5.1, which does satisfy (3.20). By what we have shown above,

$$|(PW^*)'(x)| \leqq C_6 \|PW^*\|_{\mathbb{R}} \int_{Q^*(\max\{1,|x|\})}^{C_5 n} ds/Q^{*[-1]}(s), \tag{5.15}$$

$P \in \mathcal{P}_n$, $n \geqq C_7$, and $|x| \leqq \eta a_n$. In view of (5.2)-(5.4), we obtain (1.15) for W for $\rho \leqq |x| \leqq \eta a_n$, any fixed $\rho > 0$. If in addition Q' is continuous at zero, then it follows from the continuity of Q and Q' that

$$|(PW)'(x)| \leqq C_{10}\{|(PW^*)'(x)| + |(PW)(x)|\},$$

$|x| \leqq \rho$, $P \in \mathcal{P}_n$, $n \geqq 1$. Applying (5.15) and (5.2) to (5.4) yields (1.15) for $|x| \leqq \rho$. □

We proceed to the following proof.

Proof of (1.14) *of Theorem* 1.3 *for* $|x| \in [\eta a_n, 2a_n]$. Let $0 < \eta < 1$ and $x \in [\eta a_n, 2a_n]$. Furthermore, let $s := x/a_n \in [\eta, 2]$ and

$$\varepsilon := \begin{cases} a_n \cdot \min\{n^{-2/3}, n^{-1}(1-s)^{-1/2}\} & \text{if } s \leqq 1, \\ a_n \cdot n^{-2/3} & \text{if } s > 1. \end{cases} \tag{5.16}$$

As usual, we estimate τ and $\max_{|t-x|=\varepsilon} \exp(nU_n(t/a_n))$ in (2.3) and (5.8). Since for n large enough, $\varepsilon \leqq a_n n^{-2/3} \leqq a_n \eta/2 \leqq x/2$, (2.4) shows that

$$\begin{aligned} \tau &= C2xQ'(2x)\{\varepsilon/(x-\varepsilon)\}^2 \\ &\leqq C_3 n n^{-4/3}/(\eta - n^{-2/3})^2 \leqq C_4 n^{-1/3}, \end{aligned}$$

by (3.6). Suppose now $|t-x| = \varepsilon$.

First, if $\operatorname{Re} t \leqq a_n$, (4.11) in Lemma 4.3 yields

$$\begin{aligned} U_n(t/a_n) &\leqq C_4 \max\{|\operatorname{Im} t/a_n|^{3/2}, |\operatorname{Im} t/a_n|(1-\operatorname{Re} t/a_n)^{1/2}\} \\ &\leqq C_4 \max\{(\varepsilon/a_n)^{3/2}, (\varepsilon/a_n)(1-\operatorname{Re} t/a_n)^{1/2}\}. \end{aligned}$$

Here, if $s \leqq 1$,

$$\begin{aligned} (1-\operatorname{Re} t/a_n)^{1/2} &\leqq (1-s)^{1/2} + |s-\operatorname{Re} t/a_n|^{1/2} \\ &\leqq (1-s)^{1/2} + (\varepsilon/a_n)^{1/2}, \end{aligned}$$

so

$$nU_n(t/a_n) \leqq 2C_4 \max\{n(\varepsilon/a_n)^{3/2}, n(\varepsilon/a_n)(1-s)^{1/2}\} \leqq 2C_4,$$

by the choice (5.16) of ε. If $s \geqq 1$,

$$(1-\operatorname{Re} t/a_n)^{1/2} \leqq (s-\operatorname{Re} t/a_n)^{1/2} \leqq (\varepsilon/a_n)^{1/2},$$

so

$$nU_n(t/a_n) \leqq C_4 n(\varepsilon/a_n)^{3/2} \leqq C_4.$$

Next, if $\operatorname{Re} t > a_n$, (4.12) and (4.13) in Lemma 4.4 yield

$$nU_n(t/a_n) \leqq C_4 n|\operatorname{Im} t/a_n|^{3/2} \leqq C_4 n(\varepsilon/a_n)^{3/2} \leqq C_4.$$

Thus

$$\max_{|t-x|=\varepsilon} \exp(nU_n(t/a_n)) \leqq C_5.$$

Substituting this last estimate in (5.8), and then that for τ in (2.3), easily yields (1.14) for $x \in [\eta a_n, 2a_n]$. Since W is even, the estimate (1.14) follows also for $x \in [-2a_n, -\eta a_n]$. □

Before proving (1.14) for the remaining range $|x| \geqq 2a_n$, we need the following proof.

Proof of Theorem 1.1. Suppose first Q' is continuous at 0. Then

$$\begin{aligned} |(P'W)(x)| &\leqq |(PW)'(x)| + |Q'(x)||(PW)(x)| \\ &\leqq C_3 \|PW\|_{\mathbb{R}} \max\left\{n/a_n, \int_{Q(1)}^{C_5 n} ds/Q^{[-1]}(s), |Q'(x)|\right\}, \end{aligned}$$

for $n \geqq C_4$, $P \in \mathcal{P}_n$, and $|x| \leqq 2a_n$, by (1.15) and (1.14) for the restricted range $|x| \in [\eta a_n, 2a_n]$. We will see that the second term in $\{\}$ is the largest, except possibly for multiplication by a constant. First, $|Q'(x)|$ is bounded above by a positive constant in any finite interval.

Next, suppose $1 \leqq |x| \leqq a_n$. If C_5 is large enough, (3.6) and the substitution $s = Q(t)$ yield

$$\begin{aligned} \int_{Q(1)}^{C_5 n} ds/Q^{[-1]}(s) &\geqq \left[\int_1^{2a_n} + \int_{2a_n}^{3a_n}\right] Q'(u)u^{-1}\,du \\ &\geqq xQ'(x)\int_{|x|}^{2|x|} u^{-2}\,du + C_6 n/a_n \int_{2a_n}^{3a_n} u^{-1}\,du \\ &= |Q'(x)|/2 + C_7 n/a_n, \end{aligned}$$

where we have used the evenness and monotonicity of $uQ'(u)$ and (3.6). Thus

$$|(P'W)(x)| \leqq C_3 \|PW\|_{\mathbb{R}} \int_{Q(1)}^{C_5 n} ds/Q^{[-1]}(s),$$

$|x| \leqq 2a_n$, $P \in \mathcal{P}_n$, $n \geqq C_6$. It is easily seen that we may replace $Q(1)$ by 1 as the lower limit in this integral, so (2.20) yields (1.6) in this case.

When $Q'(x)$ is not continuous at 0, we replace W by W^* of Lemma 5.1, exactly as in the proof of (1.15) of Theorem 1.3, and use (5.3) and (5.4). □

It remains to complete the following proof.

Proof of (1.14) *of Theorem* 1.3 *for* $|x| \in [2a_n, \infty)$. Now if $P \in \mathcal{P}_n$ and $|x| \geqq 2a_n$, then (2.19) yields

$$\begin{aligned} |(PW)'(x)| &\leqq |(P'W)(x)| + |Q'(x)||(PW)(x)| \\ &\leqq \exp(nU_n(x/a_n))\{\|P'W\|_{\mathbb{R}} + |Q'(x)|\,\|PW\|_{\mathbb{R}}\} \\ &\leqq C_3 \|PW\|_{\mathbb{R}} \exp(nU_n(x/a_n))\{n + |x|^{B-1}\}, \end{aligned}$$

by (1.6) and the fact that

$$Q^{[-1]}(s) \geqq Q^{[-1]}(1), \qquad s \in (1, \infty),$$

as well as by (3.2). Next,

$$\begin{aligned} &\exp(nU_n(x/a_n))\{n + |x|^{B-1}\} \\ &\quad \leqq \exp(-nC \log(|x|/a_n) + \log n + (B-1)\log|x|) \qquad \text{(by (2.18))} \\ &\quad = \exp(\{\log(|x|/a_n)\}\{-nC + B - 1\} + \log n + (B-1)\log a_n) \\ &\quad \leqq \exp(-nC_3 + C_4 \log n), \end{aligned}$$

by (3.5). This last right-hand side is certainly bounded above by $n^{-2/3}$ for n large enough. □

REFERENCES

[1] Z. DITZIAN AND V. TOTIK, *Moduli of Smoothness*, Springer Ser. Comput. Appl. Math., Vol. 9, Springer-Verlag, Berlin, New York, 1987.

[2] M. M. DZRBASJAN, *Some questions in the theory of weighted polynomial approximation on the entire real axis*, Mat. Sb., 36 (1955), pp. 355-440. (In Russian.)

[3] G. FREUD, *Markov-Bernstein type inequalities in* $L_p(-\infty, \infty)$, in Approximation Theory II, G. G. Lorentz et al. eds., Academic Press, New York, 1976, pp. 369-377.

[4] ———, *On Markov-Bernstein type inequalities and their applications*, J. Approx. Theory, 19 (1977), pp. 22-37.

[5] A. L. LEVIN AND D. S. LUBINSKY, *Canonical products and the weights* $\exp(-|x|^\alpha)$, $\alpha > 1$, *with applications*, J. Approx. Theory, 49 (1987), pp. 149-169.

[6] ———, *Weights on the real line that admit good relative polynomial approximation, with applications*, J. Approx. Theory, 49 (1987), pp. 170-195.

[7] D. S. LUBINSKY, *L_∞ Markov and Bernstein Inequalities for Erdös Weights*, J. Approx. Theory, to appear.

[8] D. S. LUBINSKY AND P. NEVAI, *Markov-Bernstein inequalities revisited*, J. Approx. Theory Appl., 3 (1987), pp. 98-119.

[9] D. S. LUBINSKY AND E. B. SAFF, *Strong Asymptotics for Extremal Polynomials Associated with Weights on* $\mathbb{R}$, Lecture Notes in Math. 1305, Springer-Verlag, Berlin, New York, 1988.

[10] H. N. MHASKAR AND E. B. SAFF, *Extremal problems for polynomials with exponential weights*, Trans. Amer. Math. Soc., 285 (1984), pp. 203-234.

[11] ———, *Where does the* Sup *norm of a weighted polynomial live?* Constr. Approx., 1 (1985), pp. 71-91.

[12] P. NEVAI, *Geza Freud, Orthogonal polynomials and Christoffel functions. A case study*, J. Approx. Theory, 48 (1986), pp. 3-167.

[13] P. NEVAI AND V. TOTIK, *Weighted polynomial inequalities*, Constr. Approx. 2 (1986), pp. 113-127.

[14] W. E. SEWELL, *Degree of Approximation by Polynomials in the Complex Domain*, Princeton University Press, Princeton, NJ, 1942.

[15] J. L. WALSH, *Interpolation and Approximation by Rational Functions in the Complex Domain*, Amer. Math. Soc. Colloq. Publ., Vol. 20, Providence, R.I., 1935, Third edition, 1960.

SIAM J. MATH. ANAL.
Vol. 21, No. 4, pp. 1083-1092, July 1990
© 1990 Society for Industrial and Applied Mathematics
015

WICK-WIGNER FUNCTIONS AND TOMOGRAPHIC METHODS*

NICOLAS LERNER†

Abstract. The ambiguity function, occurring in radar problems, is the total Fourier transform of the Wigner function, the latter being linked with Weyl quantization of pseudodifferential operators. This remark leads to a microlocal approach for tomographic problems. First, an exact inversion formula for a distribution of objects is proved, using the symplectic invariance of the Weyl quantization. Second, asymptotic methods of microlocal analysis give an approximate inversion formula, involving fewer oscillatory integrals than in the first case.

Key words. tomography, radar function, pseudodifferential operators

AMS(MOS) subject classifications. 35R30, 35S99, 47G05

Introduction. Tomographic reconstruction problems are a topic of growing interest on a theoretical point of view as well as a numerical one [4]-[7]. The main goal of the present paper is to show that a large array of methods coming from partial differential equations can be applied to these problems.

First, we remark that the ambiguity function is the Fourier transform of the Wigner function (see, e.g., [9]), the latter being linked with Weyl quantization. The main consequence of this fact is the symplectic invariance of the ambiguity function, which appears as a consequence of the Segal formula.

Second, we remark that there is a natural way to associate a pseudodifferential operator to a distribution of objects. Let us say here briefly that a distribution of objects will have a Fourier transform in L^∞, and that Fourier transform will be the symbol of the considered pseudodifferential operator. Moreover, the tomographic problem is thus reduced to a "generic" one, namely recovering a symbol by using Wigner's functions with knowledge of his dot products.

As a consequence of these two opening remarks it is not surprising to find on our way the uncertainty principle: the phase space is symplectically incompressible, e.g., it is hopeless to localize a particle in a box of respective sides Δp, Δq unless the product $\Delta p \Delta q$ is bounded from below by a fixed constant. For instance, if we assume the distribution $\hbar$-fuzzy, i.e., nonclearly resolved but supported in boxes of symplectic volume n, then its Fourier transform will be in a very simple asymptotic class of pseudodifferential operators; it is then possible to apply classical methods to handle the tomographic problem. In particular, the wave packets introduced by Cordoba and Fefferman [3] (see also Unterberger [11]) will be helpful.

On the other hand, we also provide one exact inversion formula linked to the structure of the symplectic group: the integration of the ambiguity function of the distribution through the orbit of the Gaussian pulse by a subgroup of the symplectic group gives exactly the initial distribution. Here we have used phase translations.

1. Definitions.

1.1. Basic notation and properties. We will consider the configuration space $E = \mathbb{R}^n_x$ and its dual space $E^* = \mathbb{R}^n_\xi$,

$$F = E \oplus E^* \tag{1.1}$$

* Received by the editors November 9, 1987; accepted for publication (in revised form) September 24, 1989. This work was supported in part by National Science Foundation grant DMS-8601755.

† Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

the phase space. The running point of F will be denoted generally by a capital letter ($X=(x,\xi)$, $Y=(y,\eta)\cdots$). The symplectic form on F is given by

$$[(x,\xi);(y,\eta)]=\langle\xi,y\rangle-\langle\eta,x\rangle, \tag{1.2}$$

where $\langle,\rangle$ are the brackets of duality E, E^*. The (affine) symplectic group is the subgroup of the affine group of F preserving the symplectic form (1.2). Note also that if $\sigma=\left(\begin{smallmatrix}0 & \mathrm{id}_{(E^*)}\\ -\mathrm{id}_{(E)} & 0\end{smallmatrix}\right)$ we have

$$[X,Y]=\langle\sigma X,Y\rangle_{F^*,F} \tag{1.3}$$

and the equation of the (linear) symplectic group is thus

$$A^*\sigma A=\sigma. \tag{1.4}$$

We recall here (without proof) Lemma 18.5.8 in [8] describing the symplectic group.

LEMMA 1.1.1. *The affine symplectic group is spanned by*

(1) $$(x,\xi)\to(x+x_0,\xi+\xi_0),$$

(2) $$(x,\xi)\to(Ax,A^{-1}\xi),$$

where A is an invertible $n\times n$ matrix.

(3) $$(x_j,\xi_j)\to(\xi_j,-x_j)$$

and other coordinates are fixed.

(4) $$(x,\xi)\to(x+S\xi,\xi),$$

where S is a symmetric matrix.

1.2. Wigner's function and ambiguity function. Let u, v be in the Schwartz space $\mathcal{S}(\mathbb{R}^n)$ and set

$$H(u,v)(x,\xi)=\int u\left(x+\frac{z}{2}\right)\bar v\left(x-\frac{z}{2}\right)e^{-\mathrm{i}z\xi}\,dz. \tag{1.5}$$

Let us consider for $(x,\xi)=X$ fixed the following operator (phase symmetry operator):

$$\left(\sigma_{X=(x,\xi)}u\right)(y)=u(2x-y)\,e^{-2\mathrm{i}\langle x-y,\xi\rangle}, \tag{1.6}$$

with the notation

$$\mathrm{i}=2\pi\sqrt{-1}. \tag{1.7}$$

We will use the following definition for the Fourier transform $Fu(\xi)=\int e^{-\mathrm{i}x\xi}u(x)\,dx$, so that $F^2=C$, $Cu(x)=u(-x)$.

LEMMA 1.2.1. (1) *The sesquilinear mapping $(u,v)\mapsto H(u,v)$ is continuous from $\mathcal{S}(\mathbb{R}^n)\times\mathcal{S}(\mathbb{R}^n)$ in $\mathcal{S}(\mathbb{R}^{2n}_{x,\xi})$.*

(2) *For each $X\in\mathbb{R}^{2n}$, σ_X is a unitary and self-adjoint operator on $L^2(\mathbb{R}^n)$.*

(3) *For u, v in $L^2(\mathbb{R}^n)$ $H(u,v)(X)=2^n(\sigma_X u,v)_{L^2(\mathbb{R}^n)}$.*

(4) *The sesquilinear Hermitian mapping*

$$H:\quad L^2(\mathbb{R}^n)\times L^2(\mathbb{R}^n)\to L^2(\mathbb{R}^{2n})$$

defined by (1.5) *is such that*

$$\|H(u,v)\|^2_{L^2(\mathbb{R}^{2n})}=\|u\|^2_{L^2(\mathbb{R}^n)}\|v\|^2_{L^2(\mathbb{R}^n)}.$$

Proof. The first three points are obvious. Let us prove (4). Let $\omega(x, z)$ be an $L^2(dx\,dz)$ function and let us consider its partial Fourier transform with respect to the second variable $\Omega(x, \xi) = F_2(\omega)(x, \xi)$. We have $\|\Omega\|_{L^2} = \|\omega\|_{L^2}$, which gives (4) with $\omega(x, z) = u(x + (z/2))\bar{v}(x - (z/2))$.

We will say that $H(u, v)$ is "the Wigner function of u and v." Note that

$$H(u) = H(u, u) \tag{1.8}$$

the "Wigner function of u." As is shown in the next section, the definition of Wigner's function is not arbitrary but linked to the Weyl quantization.

We define, for u, v in $L^2(\mathbb{R}^n)$, the ambiguity function $A(u, v)$ as

$$A(u, v) = F(H(u, v)), \tag{1.9}$$

that is, the (total) Fourier transform of the Wigner function. We get easily the following *classical* integral formula for $A(u, v)$

$$A(u, v)(\xi, x) = \int u\left(z - \frac{x}{2}\right)\bar{v}\left(z + \frac{x}{2}\right) e^{-i z\xi}\, dz. \tag{1.10}$$

(It seems natural to reverse the order of the variables: (x, ξ) on the phase space (ξ, x) on its dual). Note that

$$A(u) = A(u, u) = FH(u), \tag{1.11}$$

the "ambiguity function" of u.

1.3. Quantization. Let a be a tempered distribution of $\mathcal{S}'(\mathbb{R}^{2n}_{x,\xi})$. Then we define the operator a^ω as a continuous operator from $\mathcal{S}(\mathbb{R}^n)$ in $\mathcal{S}'(\mathbb{R}^n)$ by

$$(a^w u, v)_{\mathcal{S}^*(\mathbb{R}^n),\mathcal{S}(\mathbb{R}^n)} = \langle a, H(u, v)\rangle_{\mathcal{S}'(\mathbb{R}^{2n}),\mathcal{S}(\mathbb{R}^{2n})}. \tag{1.12}$$

Note that this formula makes sense because of Lemma 1.2.1 (1). The operator a^w is the operator with symbol a by the Weyl quantization rule. We get (when it makes sense) the usual formula

$$(a^w u)(x) = \int\int a\left(\frac{x+y}{2}, \xi\right) e^{i\langle x-y,\xi\rangle} u(y)\, dy\, d\xi. \tag{1.13}$$

LEMMA 1.3.1. *Let a be a tempered distribution of $\mathcal{S}'(\mathbb{R}^{2n}_{x,\xi})$. We have*

(1) $$a^w = \int_{\mathbb{R}^{2n}} a(X) 2^n \sigma_X\, dX,$$

where σ_X is defined in (1.6). *Moreover,*

(2) $$a^w = \int_{\mathbb{R}^{2n}} F(a)(\Xi)\, e^{i\Xi.M}\, d\Xi,$$

where $\Xi.M$ is the self-adjoint operator

$$\hat{x}.x + \hat{\xi}.D_x = \sum_{j=1}^{n} \hat{x}_j x_j + \hat{\xi}_j D_{x_j},$$

where x_j stands for the multiplication by x_j and $(\hat{x}, \hat{\xi}) = \Xi$. Explicitly we get

(3) $$(a^w u)(y) = \int_{\mathbb{R}^{2n}} F(a)(\hat{x}, \hat{\xi}) \left[\left(\exp i\left(\sum_{j=1}^{n} \hat{x}_j x_j + \hat{\xi}_j D_{x_j}\right)\right) u\right](y) \times d\hat{x}_1 \cdots d\hat{x}_n\, d\hat{\xi}_1 \cdots d\hat{\xi}_n,$$

where

$$\left(\exp \mathring{\imath}\left(\sum_{j=1}^{n} \hat{x}_j x_j + \hat{\xi}_j D_{x_j}\right) u\right)(y) = \int e^{\mathring{\imath}(y-z)\cdot\eta} \exp\left(\mathring{\imath} \sum_{j=1}^{n} \hat{x}_j \cdot \frac{y_j+z_j}{z} + \hat{\xi}_j \eta_j\right) u(z)\, dz\, d\eta.$$

Proof. Point (1) is classical and is easily deduced from (1.12) and Lemma 1.2.1(3). Point (2) (the original formulation of the Weyl quantization) is a consequence of (1) by application of the Plancherel formula.

LEMMA 1.3.2. (1) *Let φ, ψ be in $L^2(\mathbb{R}^n)$. Then $M(\varphi,\psi)(x,\xi)$ is the Weyl symbol of the operator (on $L^2(\mathbb{R}^n)$):*

$$u \mapsto (u,\psi)_{L^2(\mathbb{R}^n)}\varphi \text{ (i.e., } M(\varphi,\psi)^w u = (u,\psi)_{L^2(\mathbb{R}^n)}\varphi).$$

(2) *We have the identity ($u, v, \varphi, \psi \in L^2(\mathbb{R}^n)$)*

$$(u,\psi)_{L^2(\mathbb{R}^n)}(\varphi, v)_{L^2(\mathbb{R}^n)} = (H(u,v), \overline{H(\varphi,\psi)})_{L^2(\mathbb{R}^{2n})}.$$

We have

(3) $\|a^2\|_{\mathcal{L}(L^2(\mathbb{R}^n))} \leqq 2^n \|a\|_{L^1(\mathbb{R}^{2n})}$,

(4) $\|a^w\|_{\mathcal{L}(L^2(\mathbb{R}^n))} \leqq \|F(a)\|_{L^1(\mathbb{R}^{2n})}$,

(5) $\|a^w\|_{HS} = \|a\|_{L^2(\mathbb{R}^{2n})}$, $\|\ \|_{\mathcal{L}(L^2(\mathbb{R}^n))}$ *standing for the bounded operator norm in $L^2(\mathbb{R}^n)$ and $\|\ \|_{HS}$ for the Hilbert-Schmidt norm.*

Proof. This lemma is classical. Point (3) (respectively, (4)) comes at once from Lemma 1.3.1(1) (respectively, (2)) and the unitary character of σ_X (respectively, $e^{\mathring{\imath}\Xi M}$). To get (5), let us first remark that the kernel of a^w is given by

$$k(x,y) = F_2(a)\left(\frac{x+y}{2}, y-x\right) \tag{1.14}$$

and $a = F_2(k \circ A)$, where $A = \left(\begin{smallmatrix} \mathrm{Id} & -\mathrm{Id}/2 \\ \mathrm{Id} & \mathrm{Id}/2 \end{smallmatrix}\right)$, that is, the integral formula

$$a(x,\xi) = \int e^{\mathring{\imath}t\xi} k\left(x - \frac{t}{2}, x + \frac{t}{2}\right) dt. \tag{1.15}$$

Then (5) follows directly from (1.14). Note that, as a consequence of (1.12) and the Hermitian character of H, (1) and (2) are equivalent. Moreover, the kernel of the operator in (1) is $k(x,y) = \varphi(x)\psi(y)$, so (1.15) gives (2).

Let us now recall the link between the "classical" quantization (coefficients of the differential operators on the left of differentiations) and the Weyl quantization. We first define a group $\{J^t\}_{t\in\mathbb{R}}$ by the formula

$$J^t = \exp \mathring{\imath} t D_x D_\xi, \tag{1.16}$$

so that

$$F(J^t a) = \exp \mathring{\imath} t\xi.x F(a)(\xi, x).$$

This group acts unitarily in $L^2(\mathbb{R}^{2n}_{x,\xi})$ and isomorphically in $\mathcal{S}'(\mathbb{R}^{2n}_{x,\xi})$. The "classical" quantization of the symbol a is

$$\mathrm{Op}\,(a)u(x) = \int a(x,\xi)\, e^{\mathring{\imath}x.\xi} \hat{u}(\xi)\, d\xi. \tag{1.17}$$

The J^t group occurs when we take adjoints by the formula

$$\mathrm{Op}\,(a)^* = \mathrm{Op}\,(J\bar{a}) \quad \text{where } J = J^1. \tag{1.18}$$

The "adjoint" quantization (the differentiations follow the multiplication) is given by

$$\text{Op}(Ja)u(x) = \int\int a(y,\xi)\, e^{i\langle x-y,\xi\rangle} u(y)\, dy\, d\xi. \tag{1.19}$$

The Weyl quantization ((1.2), (1.13), Lemma 1.3.1) is

$$a^w = \text{Op}(J^{1/2}a) \tag{1.20}$$

which yields in particular

$$(a^w)^* = (\bar{a})^*, \tag{1.21}$$

one of the nice features of the Weyl quantization rule.

Note also that we have ($t \in \mathbb{R}$)

$$\text{Op}(J^t a)u(x) = \int\int a((1-t)x+ty,\xi)\, e^{i\langle x-y,\xi\rangle} u(y)\, dy\, d\xi. \tag{1.22}$$

1.4. Segal formula. We recall briefly here the symplectic invariance of the Weyl calculus. There exists a group Mp (n)—the metaplectic group-subgroup of the unitary operators on $L^2(\mathbb{R}^n)$—that is a twofold covering of the symplectic group Sp (n). (Note that the π_1 group of both Mp (n) and Sp (n) is $\mathbb{Z}$) such that

$$\begin{array}{c} \text{Mp}(n) \\ \Big\downarrow \pi \\ \text{Sp}(n), \end{array}$$

where π is a homomorphism twofold covering.

If $\chi = \pi(M)$, then we have

$$H(Mu, Mv) = H(u,v) \circ \chi^{-1}, \tag{1.23}$$

where u, v are in $L^2(\mathbb{R}^n)$ and H is their Wigner function.

As an immediate consequence of (1.23) we get

$$A(Mu, Mv) = A(u,v) \circ {}^t\chi, \tag{1.24}$$

where A is the ambiguity function given by (1.9), (1.10).

In other words, if χ is an (affine) symplectic map there exists a unitary transformation M of $L^2(\mathbb{R}^n)$, uniquely determined apart from a constant factor of modulus 1 such that (1.23) is true. M is an ordinary unitary representation of Mp (n) [12]. According to Lemma 1.1.1 we need only give the expression of M for the generators of the symplectic group.

LEMMA 1.4.1. *Formula* (1.23) *is true for a symplectic mapping* χ *and a unitary operator on* $L^2(\mathbb{R}^n)$ M *as follows.*

(1) *If* $\chi(x,\xi) = (x+x_0, \xi+\xi_0)$, $M = \tau_{x_0,\xi_0}$ *with* $(\tau_{x_0,\xi_0}u)(x) = u(x-x_0)\, e^{i\langle x-(x_0/2),\xi_0\rangle}$. *Setting* $X_0 = (x_0,\xi_0)$, $Y_0 = (y_0,\eta_0)$ *we have* $\tau_{X_0}\tau_{Y_0} = e^{i/2[X_0,Y_0]}\tau_{X_0+Y_0}$, $\tau_{X_0}^* = \tau_{-X_0} = \tau_{X_0}^{-1}$, *where* $[,]$ *is the symplectic form in* (1.2).

(2) *If* $\chi(x,\xi) = (Ax^t, A^{-1}\xi)$, $A \in GL(n,\mathbb{R})$, $(Mu)(x) = |\det A|^{-1/2}u(A^{-1}x)$. *In particular, if* u *is even or odd then* $H(u)$ *is even.*

(3) *If* χ *maps* (x_j,ξ_j) *on* $(\xi_j,-x_j)$ *and leaves the other coordinates fixed,* M *will be the partial Fourier transform with respect to* x_j.

(4) *If* $\chi(x,\xi) = (x,\xi+Sx)$, S *an* $n\times n$ *symmetric matrix,* M *is defined by* $(Mu)(x) = e^{i/2\langle Sx,x\rangle}u(x)$.

This lemma is classical and a proof can be found in [8, Thm. 18.5.9]. Note that the unitary operators involved are also isomorphisms of $\mathcal{S}(\mathbb{R}^n)$.

Using (1.23), (1.12), we obtain the following lemma.

LEMMA 1.4.2. *Let a be a tempered distribution of $\mathcal{S}'(\mathbb{R}^{2n})$, χ a symplectic affine mapping, M in the fiber of χ. Then*

$$(a \circ \chi)^w = M^* a^w M.$$

1.5. Wick's function. It is a well-known fact that Wigner's function $H(u)(x, \xi)$ is not always nonnegative (in spite of the equality $\int H(u, u)(x, \xi)\, dx\, d\xi = \|u\|^2_{L^2(\mathbb{R}^n)}$, a consequence of (1.12)). Nevertheless it has been pointed out by de Bruijn [1], Cordoba and Fefferman [3], and Unterberger [11] that a Gaussian regularization of the Wigner function *is* nonnegative.

LEMMA 1.5.1. *Let $u, \varphi \in L^2(\mathbb{R}^n)$ with φ even or odd, $\|\varphi\|_{L^2} = 1$. Let $a \in L^1(\mathbb{R}^{2n})$. We have*

(1) $(H(u) * H(\varphi))(Y) = |(u, \tau_Y \varphi)_{L^2(\mathbb{R}^N)}|^2$, $Y \in \mathbb{R}^{2n}$ ($H(u)$ defined in (1.8), τ_Y *in Lemma* 1.4.1(1)) *and*

(2) $((a * H(\varphi))^w u, u) = \int a(Y) |(u, \tau_Y \varphi)|^2\, dY$.

Moreover,

(3) *If* $\varphi_0(x) = 2^{n/4} e^{-\pi|x|^2}$, $\varphi_j(x) = 2^{n/4} 2\pi^{1/2} x_j e^{-\pi|x|^2}$, $1 \leqq j \leqq n$, *we get*

$$H(\varphi_0)(x, \xi) = 2^n e^{-2\pi(|x|^2+|\xi|^2)},$$

$$H(\varphi_j)(x, \xi) = 4\pi 2^n e^{-2\pi(|x|^2+|\xi|^2)} \left(x_j^2 + \xi_j^2 - \frac{1}{4\pi} \right).$$

Proof. We have

$$(H(u) * H(\varphi))(Y) = \langle H(u), H(\varphi)(Y - \cdot) \rangle,$$

and by (2) and (1) of Lemma 1.4.1 we get

$$H(\varphi)(Y - X) = H(\varphi)(X - Y) = H(\tau_Y \varphi)(X).$$

We obtain part (1) by applying Lemma 1.3.2 (1), which gives $\langle H(u), H(\tau_Y \varphi) \rangle = |(u, \tau_Y \varphi)|^2$. Part (2) is a consequence of (1) through the definition of a^w in (1.12). The elementary computations in (3) are left to the reader.

Remark 1.5.2. When $\|\varphi\|_{L^2(\mathbb{R}^n)} = 1$, then from (1.12) we get $\iint H(\varphi)(x, \xi)\, dx\, d\xi = 1$. If, moreover, φ is even or odd, $H(\varphi)$ is even (Lemma 1.4.1 (2)) and the difference $a - a * H(\varphi)$ involves only the second derivative of a. This remark leads to a proof of Gårding's inequality with one derivative (see, e.g., Theorem 18.1.14 in [8]): Let us note that the quantization

$$a \to (a * H(\varphi))^w \tag{1.25}$$

is nonnegative (i.e., gives a nonnegative operator for a nonnegative symbol). Note that the constant function 1 is actually quantified by the identity: from (1.12), we have

$$\int_{\mathbb{R}^{2n}} H(\varphi)(X)\, dX = \|\varphi\|^2_{L^2(\mathbb{R}^n)} = 1,$$

and in particular,

$$\|u\|^2_{L^2(\mathbb{R}^n)} = \int_{\mathbb{R}^{2n}} |(u, \tau_Y \varphi)|^2\, dY.$$

We will define as the Wick function (see, e.g., [10]) of u (with respect to φ)

$$W_\varphi(u) = H(u) * H(\varphi). \tag{1.26}$$

When $\varphi = \varphi_0$ we will use only the Wick function of u,

$$W(u) = H(u) * H(\varphi_0) \tag{1.27}$$

(cf. Lemma 1.5.1 (3)).

2. Related tomographic problems.

2.1. Preliminary remarks. Let us consider a function $D(y_1, y_2)$, $y_1, y_2 \in \mathbb{R}^n$ (e.g., a distribution of objects, y_2 being the position and y_1 the velocity). If we choose as new variables $x = (2_{y2}/c)$, $\xi = (2_{\nu_0 y_1}/c)$, where ν_0 is a fixed frequency and c a parameter (speed of the electromagnetic wave) we obtain that x is running in $(\mathbb{R}_{\text{time}})^n$ and ξ in its dual space $(\mathbb{R}_{\text{frequency}})^n$. This describes how it is possible to associate a new function $d(\xi, x)$ to a distribution of objects $D(y_1, y_2)$; d is defined on a phase space (symplectic space) while D is not. Let $u \in L^2(\mathbb{R}^n)$ (a "pulse"). We consider

$$\mathscr{A}(u)(\xi, x) = \int\int d(\eta, y) A(u)(\eta - \xi, y - x)\, e^{i/2(\xi-\eta)(y+x)}\, dy\, d\eta, \tag{2.1}$$

where $A(u)$ is the ambiguity function of u (1.10). The problem at hand is recovering d from $\mathscr{A}(u)$ (or its absolute value) for some choices of the pulse u. We can note in particular that if d corresponds to a distribution of objects clearly resolved into N point clusters with reflecting intensities m_j we obtain, at least formally,

$$\mathscr{A}(u)(\xi, x) = \sum_{1 \leqq j \leqq N} m_j A(u)(\xi_j - \xi, x_j - x)\, e^{i/2(\xi-\xi_j)(x_j+x)}. \tag{2.2}$$

In particular, if $u(x) = 2^{-n/4}\, e^{-\pi|x|^2}$, we have (Lemma 1.5.1) $H(u)(x, \xi) = e^{-2\pi(|x|^2+|\xi|^2)}$, and thus

$$A(u)(\xi, x) = e^{-\pi/2(|x|^2+|\xi|^2)} 2^{-n}. \tag{2.3}$$

For this pulse, from (2.2) we get

$$\mathscr{A}(u)(\xi, x) = \sum_{j=1}^{N} m_j\, e^{i/2(\xi-\xi_j)(x+x_j)}\, e^{-(\pi/2)(|x-x_j|^2+|\xi-\xi_j|^2)} 2^{-n}. \tag{2.4}$$

In particular, if $N = 1$ the unique maximum for $\mathscr{A}(u)$ is at (x_1, ξ_1). Let us note that the function given by (2.3) is as close as possible of a Dirac function on the phase space. The uncertainty principle will allow also

$$A(u_\lambda)(\xi, x) = e^{-\pi/2(\lambda^2|x|^2+\lambda^{-2}|\xi|^2)} 2^{-n}, \tag{2.5}$$

that is (for large λ) a very broad Gaussian in ξ multiplied by a very narrow Gaussian in x both of L'-norm 1, namely

$$A(u_\lambda)(\xi, x) = e^{-\pi/2\lambda^2|x|^2} \lambda^n 2^{-n/2}\, e^{-\pi/2\lambda^{-2}|\xi|^2} \lambda^{-n} 2^{-n/2}.$$

From (1.24) and Lemma 1.4.1 we can choose $u_\lambda(x) = 2^{-n/4} \lambda^{n/2}\, e^{-\pi\lambda^2|x|^2}$.

Moreover, if we assume (as is done in some radar problems) that we measure

$$\sum_{j=1}^{N} m_j^2 |A(u)(\xi - \xi_j, x - x_j)|^2 \tag{2.6}$$

we get

$$\sum_{j=1}^{N} m_j^2 2^{-2n}\, e^{-\pi\{|x-x_j|^2+|\xi-\xi_1|^2\}}. \tag{2.7}$$

In particular, if we assume that the reflection intensities m_j are of the same order of magnitude and

$$\inf_{j\neq k}|X_j-X_k|^2=\inf_{j\neq k}|x_j-x_k|^2+|\xi_j-\xi_k|^2\gg 1,$$

we get that the function (2.7) takes N absolute maxima located near $X_j=(x_j,\xi_j)$.

So, with the two additional assumptions that the reflecting intensities are equal and that the objects are "far" from each other or if they are close this relative velocity is "large," we obtain an analogous statement than for one object: it is enough to display the "brightness pattern" and to check how many absolute maxima we have and where.

From (2.1) we have, setting $\Xi=(\xi,x)$, $N=(\eta,y)$

$$\exp-\frac{\mathring{\imath}}{4}\langle B\Xi,\Xi\rangle\mathscr{A}(u)(\Xi)=\int d(N)A(u)(N-\Xi)\,e^{-\mathring{\imath}/2([\Xi,N]+1/2\langle BN,N\rangle)}\,dN, \tag{2.8}$$

where $[\Xi,N)=-\xi y+\eta x$ the dual form of (1.3), and B the mapping from $E_\xi^*\oplus E_x\to E_x\oplus E_j^*$ given by $\left(\begin{smallmatrix}0 & \mathrm{id}_{(E^*)}\\ \mathrm{id}_{(E)} & 0\end{smallmatrix}\right)$, $\langle\cdot,\cdot\rangle$ standing for the bracket of duality $E^*\oplus E$, $E\oplus E^*$.

Note that (2.8) could be written also as

$$\mathscr{A}(u)(\Xi)=\int d(N+\Xi)A(u)(N)\,e^{-\mathring{\imath}/4\langle BN,N\rangle}\,e^{-\mathring{\imath}/2\langle\Lambda N,\Xi\rangle}\,dN, \tag{2.8$'$}$$

where $\Lambda: E_\xi^*\oplus E_x\to E_x\oplus E_\xi^*$ is given by

$$\lambda=\begin{pmatrix}0 & 0\\ 2 & 0\end{pmatrix}.$$

We have that (cf. (1.9)) for $\Xi=0$

$$\mathscr{A}(u)(0)=\int H(u)(X)\,\check{d}(N)\,e^{\mathring{\imath}XN}\,e^{-\mathring{\imath}/4\langle BN,N\rangle}\,dN\,dX,\qquad \check{d}(N)=d(-N)$$

and thus from (1.16), (1.17), (1.20), and (1.12)

$$\begin{aligned}\mathscr{A}(u)(0)&=(J^{-1/2}(F(d))^{\mathrm{Weyl}}u,u),\\ \mathscr{A}(u)(0)&=((F(d))^{\mathrm{usual}}u,u),\end{aligned} \tag{2.9}$$

that is, $\mathscr{A}(u)(0)$ is $(0p(a)u,u)$, where a is the (total) Fourier transform of d.

A reasonable assumption of regularity for $F(d)$ would be $F(d)$ in $L^\infty(\mathbb{R}^{2n})$ as well as his derivatives so that (2.9) would make sense for $u\in L^2(\mathbb{R}^n)$ (see Calderon and Vaillancourt [2]).

2.2. Uncertainty principle. Let us assume that the distribution d is not clearly resolved but "fuzzy" with an h-uncertainty, e.g.,

$$d(\xi,x)=\sum_{j=1}^N m_j\exp-\pi\left(\frac{|x-x_j|^2}{\lambda^2}+\frac{|\xi-\xi_j|^2}{\mu^2}\right)(\lambda\mu)^{-n} \tag{2.10}$$

($\alpha_i\in\mathbb{R}^n$, α, β positive numbers) with

$$\lambda\mu=h \tag{2.10$'$}$$

a positive "small" parameter.

Then we get

$$a(x,\xi)=F(d)(x,\xi)=\sum_{j=1}^N m_j\,e^{-\mathring{\imath}(x_j\xi+\xi_jx)}\exp-\pi(\mu^2|x|^2+\lambda^2|\xi|^2),$$

and in particular we get

$$|(D_\xi^\alpha D_x^\beta a)(x,\xi)|\leqq C_{\alpha\beta}\max_{j=1,\cdots,N}m_j\mu^{|\beta|}\lambda^{|\alpha|}, \tag{2.11}$$

where the $C_{\alpha,\beta}$ depend only on $|\alpha|, |\beta|, |x_j|$, and $|\xi_j|$. So $a \in S(m, \mu^2\, dx^2 + \lambda^2\, d\xi^2)$ (cf. Definition 18.4.2 in [8]).

THEOREM 2.2.1. *Assuming that d is given by* (2.10), $a = \hat{d}$,

$$\mathscr{A}(\tau_X M_\theta \varphi_0)(0) = a(X) + O(h), \tag{2.12}$$

where $\mathscr{A}(u)$ *is given in* (2.1), φ_0 *is as in Lemma* 1.5.1, τ_X, *is as in Lemma* 1.4.1 (1), *and* $(M_\theta u)(x) = \theta^{-n/2} u(\theta^{-1}x)$, $\theta = \mu^{-1/2}\lambda^{1/2}$.

Proof. We have from (2.9)

$$\mathscr{A}(\tau_X M_\theta \varphi_0)(0) = (a(x, D)\tau_X M_\theta \varphi_0, \tau_X M_\theta \varphi_0),$$

where $a(x, D) = \mathrm{Op}(a)$ given by (1.17). Because of the symbolic calculus given by (2.11) we can replace, modulo $O(h)$, $a(x, D)$ by a^w in the previous formula. Thus we get from (1.12)

$$\mathscr{A}(\tau_X M_\theta \varphi_0)(0) = \int_{\mathbb{R}^{2n}} a(Y) H(\tau_X M_\theta \varphi_0)(Y)\, dY + O(h).$$

Now from (1.23) comes

$$\mathscr{A}(\tau_X M_\theta \varphi_0)(0) = \int_{\mathbb{R}^{2n}} a(Y) H(M_\theta \varphi_0)(Y - X)\, dY + O(h),$$

that is, from (1.23) and Lemma 1.5.1(3),

$$\mathscr{A}(\tau_X M_\theta \varphi_0)(0) = \int a(Y) 2^n\, e^{-2\pi\Gamma(X-Y)}\, dY + O(h)$$

with $\Gamma(t \oplus \tau) = \theta^{-2}|t|^2 + \theta^2|\tau|^2$ and $\theta = \lambda^{1/2}\mu^{-1/2}$. By a Taylor expansion, we get

$$\mathscr{A}(\tau_X M_\theta \varphi_0)(0) = a(X) + O(h) + \int\int_0^1 (1-t) a''(X + t(Y - X)) \cdot (Y - X)^2 2^n\, e^{-2\pi\Gamma(X-Y)}\, dY\, dt.$$

But from (2.11) and (2.10) we have

$$|a''(X) T^2| \leqq Ch\Gamma(T),$$

which yields the result.

Remark 2.2.2(a). If $\lambda = \mu = h^{1/2}$, that is, if we assume that the uncertainty on range and velocity are of the same order of magnitude then $M_\theta = \mathrm{Id}$ and $a(X) \equiv \mathscr{A}(\tau_X \varphi_0)$.

(b) The same proof could be used for

$$d(\xi, x) = \sum_{j=1}^{N} m_j \exp - \pi\gamma(\Xi - \Xi_j)|\gamma|^{1/2},$$

where γ is a quadratic (positive definite) form on the (dual) phase space such that $\gamma^\sigma = h^2\gamma$, γ^σ standing for the dual metric of γ with respect to the symplectic form σ ((1.2)).

2.3. A new inverse formula. From (2.8), (1.23), and (1.9) we get ($\varphi \in \mathscr{S}(\mathbb{R}^n)$)

$$\mathscr{A}(\tau_Y \varphi)(\Xi) = \int\int d(N) H(\varphi)(X - Y)\, e^{-\mathrm{i}X(N-\Xi)} \cdot e^{-\mathrm{i}/2([\Xi, N] + 1/2\langle BN, N\rangle)}\, e^{\mathrm{i}/4\langle B\Xi, \Xi\rangle}\, dX\, dN.$$

In particular, if $\varphi = \varphi_0$ in Lemma 1.5.1 (3), we obtain

$$\mathscr{A}(\tau_Y\varphi_0)(\Xi) = \int\int d(N)2^n\, e^{-2\pi|X-Y|^2}\, e^{-i\hat{}X(N-\Xi)} \cdot e^{-i\hat{}/2([\Xi,N]+1/2\langle BN,N\rangle)}\, e^{i\hat{}/4\langle B\Xi,\Xi\rangle}\, dX\, dN,$$

and thus

$$\mathscr{A}(\tau_Y\varphi_0)(\Xi) = \int d(N)\, e^{-i\hat{}Y(N-\Xi)}\, e^{-\pi/2|N-\Xi|^2} \cdot e^{-i\hat{}/2([\Xi,N]+1/2\langle BN,N\rangle)}\, e^{i\hat{}/4\langle B\Xi,\Xi\rangle}\, dN.$$

Thus we get the following exact inverse formula, integrating the ambiguity function along the orbit of the Gaussian pulse through a subgroup of the metaplectic group (here the phase translations).

THEOREM 2.3.1. *With φ_0 given in Lemma* 1.5.1 (3), *τ_Y in Lemma* 1.4.1 (1), *and $\mathscr{A}$ by* (2.1), *we have*

$$\int_{\mathbb{R}^{2n}} \mathscr{A}(\tau_Y\varphi_0)(\Xi)\, dY = d(\Xi). \tag{2.13}$$

REFERENCES

[1] N. G. DE BRUIJN, *Uncertainty principles in Fourier analysis*, in Proc. Symposium on Inequalities, Academic Press, New York, 1987, pp. 57–71.

[2] A. P. CALDERON AND R. VAILLANCOURT, *On the boundedness of pseudodifferential operators*, J. Math. Soc. Japan, 23 (1972), pp. 374–378.

[3] A. CORDOBA AND C. FEFFERMAN, *Wave packets and Fourier integral operators*, Comm. Partial Differential Equations, 3 (1978), pp. 979–1005.

[4] Y. DAS AND W. BOERNER, *On radar target shape estimation using algorithms for reconstruction from projections*, IEEE Trans. Antennas and Propagation, 26, pp. 274–279.

[5] M. DAVISON AND F. A. GRUNBAUM, *Tomographic reconstruction with arbitrary directions*, Comm. Pure Appl. Math, 34 (1981), pp. 77–120.

[6] E. FEIG AND F. GREENLEAF, *Inversion of an integral transform associated with tomography in radar detection*, IBM RC 10900 (#48962), 1984.

[7] ———, *Tomographic methods in range-Doppler radar*, Inverse Problems, 2 (1986), pp. 185–195.

[8] L. HÖRMANDER, *The Analysis of Linear Partial Differential Operators*, Vol. III, Springer-Verlag, Berlin, New York, 1983.

[9] W. SCHEMM, *Harmonic analysis on the Heisenberg nilpotent Lie group*, Res. Notes in Math., 147, Pitman, Boston, 1986.

[10] M. A. SHUBIN, *Pseudo-differential Operators and Spectral Theory*, Springer-Verlag, Berlin, New York, 1985.

[11] A. UNTERBERGER, *Oscillateur harmonique et opérateurs pseudo-différentiels*, Ann. Inst. Fourier (Grenoble), 29 (1979), pp. 201–221.

[12] A. WEIL, *Sur certains groupes d'opérateurs unitaires*, Acta. Math., 111 (1964), pp. 143–211.

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1093-1117, September 1990
© 1990 Society for Industrial and Applied Mathematics

NONHOMOGENEOUS VISCOUS INCOMPRESSIBLE FLUIDS: EXISTENCE OF VELOCITY, DENSITY, AND PRESSURE*

JACQUES SIMON†

Abstract. The flow of a nonhomogeneous viscous incompressible fluid that is known at an initial time $t=0$ is considered. Such a flow is described by partial differential equations for the velocity u, the density ρ, and the pressure p, with boundary and initial conditions.

The existence of a global (in time) solution u, ρ, p for which ρu satisfies a weak initial condition is proved. For this solution u and ρu are not necessarily t-continuous, and $u(0)$ and $(\rho u)(0)$ are not defined. The initial density ρ_0 is not required to have a positive lower bound.

When u_0, f, and Ω are regular, the solution is regular up to some time T_*. For this solution, ρu is t-continuous up to T_* and satisfies an initial condition that is intermediate between the weak and the strong ones.

If in addition ρ_0 is not too small, but possibly zero at some points, then u is t-continuous at $t=0$ and satisfies the strong initial conditions $(\rho u)(0)=\rho_0 u_0$, $u(0)=u_0$, and u, ρ, p is a global strong solution. In space dimension 2 the solution is regular for all t if ρ_0 is bounded from below.

Key words. Navier-Stokes, existence, strong solutions

AMS(MOS) subject classification. 35Q10

Introduction.

Model. We consider the flow of a fluid that is viscous, incompressible, and nonhomogeneous, that is, with a variable density. It is considered in a domain $\Omega \subset \mathbb{R}^3$ with boundary Γ, during a time interval $[0, T]$. The velocity u, the pressure p, and the density ρ satisfy the following equations in $\Omega \times]0, T[$:

$$\frac{\partial \rho u}{\partial t} + \nabla \cdot (u \rho u) - \mu \Delta u = \rho f - \nabla p,$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho u) = 0,$$

$$\nabla \cdot u = 0$$

coupled with the boundary and the initial conditions

$$u = 0 \quad \text{on } \Gamma \times]0, T[$$

$$\rho|_{t=0} = \rho_0, \quad \rho u|_{t=0} = \rho_0 u_0 \quad \text{in } \Omega.$$

Solution of a reduced model. A reduced model is obtain by eliminating the pressure. More precisely, the first equation and the last initial condition are replaced by the following variational equation:

$$\int_{\Omega \times]0,T[} \left(-\rho u \cdot \frac{\partial \varphi}{\partial t} - \mu \rho u \cdot \nabla \varphi + \mu \nabla u \cdot \nabla \varphi - \rho f \cdot \varphi \right) dx\, dt = \int_\Omega \rho_0 u_0 \cdot \varphi(0)\, dx$$

$$\forall \varphi \in C^1([0, T]; (H_0^1(\Omega))^3) \quad \text{such that } \nabla \cdot \varphi = 0 \text{ and } \varphi(T) = 0.$$

A solution ρ, u of this model has been obtained by Antonzev and Kajikov [2] and by Kajikov [5]. They consider $u_0 \in H$, $\rho_0 \in L^\infty(\Omega)$ and they suppose that ρ_0 has a lower bound >0. For a detailed exposition we refer the reader to Lions [9], [10]; the definition of the space H will be recalled in § 2.

* Received by the editors March 1, 1989; accepted for publication (in revised form) October 24, 1989.
† Département de Mathématiques Appliquées, Université Blaise Pascal (Clermont-Ferrand), 63177 AUBIERE Cedex, France.

This result was extended without restriction on ρ_0 by Simon [13]. The local existence, that is, for $T = T_*$, where $T_* > 0$ depends on the data, of a more regular solution of this reduced model has been obtained by Kim [7], for more regular u_0, f, and Ω. In all these works the t-continuity of u, or of ρu, was not proved; therefore the initial condition on ρu included in the variational equation was satisfied in a very weak sense.

Solution of a weak model. In the present work we obtain (Theorem 9) a solution u, ρ, p of the exact model including the pressure, with the exception of the initial condition on ρu that is replaced by the following weak initial condition:

$$\left(\int_\Omega \rho u \cdot v \, dx \right) \Big|_{t=0} = \int_\Omega \rho_0 u_0 \cdot v \, dx$$

for a certain class of test functions v.

The other equations and conditions are satisfied in distribution and in trace sense. In addition the above integral is t-continuous for each v, although ρu is not necessarily t-continuous. Moreover, the pair u, ρ satisfies the variational equation of the reduced model.

We do not require ρ_0 to have a positive lower bound, and we only assume f to be in $L^1(0, T; (L^2(\Omega))^3)$ instead of $L^2(0, T; H)$ as is done in the above-quoted works (we do not require f to be divergence free). Our solution is global, that is, it exists for any given T.

Solution of the exact model. The local existence of a solution u, ρ, p of all the exact equations has been proved by Ladyzenskaya and Solonnikov [18]: with regular data and ρ_0 bounded from below they obtain for $T = T_*$, where T_* depends on the data, a regular solution. In addition they prove the global existence for small enough data and in the two-dimensional case (for all data), and the uniqueness, in their class, of regular solutions. Similar results have been obtained by Okamoto [20] with different assumptions on the regularity of the data.

Here we prove (Theorem 14(ii)) the global existence of a solution u, ρ, p for all data. Our data are less regular than in [18], and are not assumed to be small; in addition ρ_0 may have some zeros. The solution exists up to T, and ρu is t-continuous up to some $T_* > 0$ (this is enough to satisfy the initial condition on ρu).

More precisely we require $u_0 \in V$, $\rho_0 \in L^\infty(\Omega)$ with $1/\rho_0 \in L^{6/5}(\Omega)$, and $f \in L^2(0, T; (L^2(\Omega))^3)$ (the space V will be precisely defined in § 2).

Solution of an intermediate model. Without the assumption on $1/\rho_0$, that is, for all $\rho_0 \in L^\infty(\Omega)$, we obtain (Theorem 14(i)) the same result with the exception of the initial condition on ρu that is replaced by

$$\rho u|_{t=0} = \rho_0 u_0 + \nabla \lambda \quad \text{where } \lambda \text{ is an unknown function.}$$

This condition is intermediate between the weak and the strong one.

1. Mathematical model. Let Ω be an open bounded subset of $\mathbb{R}^3$ with boundary Γ, and let $T > 0$. The motion of the fluid is described by its velocity $u = (u_1, u_2, u_3)$, by its density ρ, and by its pressure p, which are functions of the point $x \in \Omega$ and of the time $t \in [0, T]$.

The equations are, in the cylinder $Q=\Omega\times]0, T[$,

$$\frac{\partial \rho u}{\partial t}+\nabla\cdot(u\rho u)-\mu\Delta u=\rho f-\nabla p, \tag{1}$$

$$\frac{\partial \rho}{\partial t}+\nabla\cdot(\rho u)=0, \tag{2}$$

$$\nabla\cdot u=0 \tag{3}$$

where f is the given external force, and $\mu>0$.

These equations are coupled with the boundary condition

$$u=0 \quad \text{on } \Gamma\times]0, T[$$

and with the initial conditions, for $t=0$,

$$\rho(0)=\rho_0, \qquad (\rho u)(0)=\rho_0 u_0.$$

Here ρ_0 and u_0 are given functions in Ω, and $u(t)$ is the function $x\to u(x, t)$. Moreover, the density must satisfy

$$\rho\geqq 0, \qquad \rho_0\geqq 0.$$

Here $\nabla=(\partial_1, \partial_2, \partial_3)$, where $\partial_i=\partial/\partial x_i$ is the gradient operator, and $\cdot$ is the scalar product in $\mathbb{R}^3$. Thus $\nabla\cdot u=\partial_1 u_1+\partial_2 u_2+\partial_3 u_3$ and $\nabla\cdot(u\rho u)=\partial_1(u_1\rho u)+\partial_2(u_2\rho u)+\partial_3(u_3\rho u)$. Also, $\Delta=\nabla\cdot\nabla$ is the Laplace operator.

Remark. Nonconservative equations. Equations (1) and (2) may be replaced by the following:

$$\rho\left(\frac{\partial u}{\partial t}+(u\cdot\nabla)u\right)-\mu\Delta u=\rho f-\nabla p, \tag{1'}$$

$$\frac{\partial \rho}{\partial t}+u\cdot\nabla\rho=0. \tag{2'}$$

Indeed (1') and (1) differ from u times (2), and (2') and (2) differ from ρ times (3). We will use (1) and (2) since they require less smoothness on ρ and u to be defined.

Remark. Nonhomogeneous fluids. Such a fluid is obtained by mixing two miscible fluids that are incompressible and that have different densities. It may also be a fluid containing a melted substance.

2. Function spaces and preliminaries. The set Ω is assumed to be open and bounded in $\mathbb{R}^3$ and to have a Lipschitz boundary Γ.

Let $\mathcal{D}(\Omega)$ be the space of C^∞ functions with compact support in Ω, and let $\mathcal{D}'(\Omega)$ be the space of distribution on Ω.

For $1\leqq r\leqq\infty$ the Sobolev spaces are defined by

$$W^{1,r}(\Omega)=\{v\in L^r(\Omega): \nabla v\in (L^r(\Omega))^3\},$$

$$W_0^{1,r}(\Omega)=\text{closure of } \mathcal{D}(\Omega) \text{ in } W^{1,r}(\Omega),$$

$$W^{-1,r}(\Omega)=\left\{v=v_0+\sum_{i=1}^{3}\partial_i v_i : v_i\in L^r(\Omega), i=0,\cdots,3\right\},$$

$$H^1(\Omega)=W^{1,2}(\Omega), \quad H_0^1(\Omega)=W_0^{1,2}(\Omega), \quad H^{-1}(\Omega)=W^{-1,2}(\Omega),$$

all these spaces being equipped with their usual norms.

Denote

$$\mathcal{V} = \{v \in (\mathcal{D}(\Omega))^3 : \nabla \cdot v = 0\},$$

V = closure of $\mathcal{V}$ in $(H^1(\Omega))^3$, equipped with the $(H^1(\Omega))^3$ norm,

H = closure of $\mathcal{V}$ in $(L^2(\Omega))^3$, equipped with the $(L^2(\Omega))^3$ norm.

Remark. Since the boundary Γ of Ω is Lipschitz we have

$$V = \{v : v \in (H^1(\Omega))^3, \nabla \cdot v = 0, v|_\Gamma = 0\},$$

$$H = \{v : v \in (L^2(\Omega))^3, \nabla \cdot v = 0, v|_\Gamma \cdot n = 0\}$$

where $v|_\Gamma$ is the trace of v on Γ and n is a normal vector field on Γ.

See Propositions 1 and 2 of [17, pp. 25, 26], or Corollary 2.5 and Theorem 2.8 of [4].

We denote by $(\ ,\ )_\Omega$ the duality product in all the function spaces on Ω. In particular,

$$(v, w)_\Omega = \int_\Omega v(x)w(x)\,dx \quad \text{if } v \in L^r(\Omega), \quad w \in L^{r'}(\Omega), \quad \frac{1}{r} + \frac{1}{r'} = 1,$$

($v(x)w(x)$ being replaced by $v(x) \cdot w(x)$ if v and w are vector valued) and

$$(v, w)_\Omega = w(v) \quad \text{if } v \in \mathcal{D}(\Omega), \quad w \in \mathcal{D}'(\Omega).$$

Let us recall a result of Tartar [17, Lemma 9, p. 30] (see also [4, Thm. 2.3, p. 2.5]).

LEMMA 1. *Let $w \in (H^{-1}(\Omega))^3$ satisfy $(v, w)_\Omega = 0$, for all $v \in \mathcal{V}$. Then there exists $q \in L^2(\Omega)$ such that $w = \nabla q$.*

Moreover q may be chosen such that the map $L : w \to q$ is linear and continuous from $(H^{-1}(\Omega))^3$ into $L^2(\Omega)$.

Let us extend this result to time-dependent distributions.

LEMMA 2. *Let $h \in \mathcal{D}'(]0, T[; (H^{-1}(\Omega))^3)$ satisfy $(h, v)_\Omega = 0$, for all $v \in \mathcal{V}$. Then there exists $g \in \mathcal{D}'(]0, T[; L^2(\Omega))$ such that $h = \nabla g$.*

Moreover, g may be chosen such that the map $h \to g$ is linear and continuous from $W^{s,r}(0, T; (H^{-1}(\Omega))^3)$ into $W^{s,r}(0, T; L^2(\Omega))$, for all $s \in \mathbb{R}$ and $1 \leqq r \leqq \infty$.

Let us recall that, E being a Banach space, $\mathcal{D}'(]0, T[; E)$ is the set of continuous linear maps from $\mathcal{D}(]0, T[)$ into E.

Proof of Lemma 2. The space

$$E = \{w \in (H^{-1}(\Omega))^3 : (w, v)_\Omega = 0, \forall v \in V\}$$

equipped with the $(H^{-1}(\Omega))^3$ norm is a Banach space.

The distribution $(h, v)_\Omega \in \mathcal{D}'(]0, T[)$ is defined by

$$((h, v)_\Omega, \varphi)_{]0,T[} = ((h, \varphi)_{]0,T[}, v)_\Omega,$$

and thus the assumption on h implies that $h \in \mathcal{D}'(]0, T[; E)$.

The map L defined in Lemma 1 is linear continuous from E into $L^2(\Omega)$; thus $Lh \in \mathcal{D}'(]0, T[; L^2(\Omega))$, and $g = Lh$ has the required properties. □

Since Ω is a bounded subset of $\mathbb{R}^3$ with Lipschitz boundary, $H^1(\Omega) \subset L^6(\Omega)$, and more generally the following properties hold.

LEMMA 3. (i) *For $1 \leqq r \leqq \infty$, $W^{1,r}(\Omega) \subset L^{r*}(\Omega)$ with continuous imbedding, where*

$$\frac{1}{r_*} = \frac{1}{r} - \frac{1}{3} \text{ if } r < 3, \quad r_* \text{ is any finite real if } r = 3, \quad r_* = \infty \text{ if } r > 3.$$

The imbedding $W^{1,r}(\Omega) \to L^s(\Omega)$ is compact if $s < r_*$.

(ii) *For* $1 \leqq s \leqq r \leqq \infty$ *the product is continuous, if* $t \geqq 1$, *in*

$$W^{1,r}(\Omega) \times W^{1,s}(\Omega) \to W^{1,t}(\Omega), \qquad \frac{1}{t} = \frac{1}{r_*} + \frac{1}{s}.$$

(iii) *For* $1 \leqq r \leqq \infty$, $1 \leqq s \leqq \infty$, *the product is continuous, if* $1/r + 1/s \leqq 1$, *in*

$$W^{1,r}(\Omega) \times W^{-1,s}(\Omega) \to W^{-1,t}(\Omega), \qquad \frac{1}{t} = \frac{1}{r_*} + \frac{1}{s}.$$

Proof. (i) For Sobolev's theorem see, for example, [1, p. 97].

(ii) Let $v \in W^{1,r}(\Omega)$, $w \in W^{1,s}(\Omega)$. There holds $vw \in L^a(\Omega)$, $1/a = 1/r_* + 1/s_*$, $\nabla vw = v\nabla w + w\nabla v \in L^b(\Omega) + L^c(\Omega)$, $1/b = 1/r_* + 1/s$, $1/c = 1/s_* + 1/r$. Thus $vw \in L^t(\Omega)$, $t = \inf(a, b, c)$. For $s \leqq r$ this gives $t = b$.

(iii) Let $v \in W^{1,r}(\Omega)$, $w \in W^{-1,s}(\Omega)$, that is, $w = w_0 + \Sigma_i \partial_i w_i$, $w_i \in L^s(\Omega)$, and $\varphi = vw_0 - \Sigma_i w_i \partial_i v + \Sigma_i \partial_i (vw_i)$.

There holds $v \in L^{r}{}_*(\Omega)$, $w_i \in L^s(\Omega)$; thus $vw_i \in L^t(\Omega)$, $1/t = 1/r_* + 1/s$, and $\partial_i(vw_i) \in W^{-1,t}(\Omega)$.

On the other hand, $w_i \partial_i v \in L^a(\Omega)$, $1/a = 1/r + 1/s$, and $vw_0 \in L^a(\Omega)$. By Sobolev's theorem $L^a(\Omega) \subset W^{-1,t}(\Omega)$, since $t \leqq a_*$. Finally, $\varphi \in W^{-1,t}(\Omega)$. Moreover, taking all the possible w_i, we obtain

$$|\varphi|_{W^{-1,t}(\Omega)} \leqq c|v|_{W^{-1,r}(\Omega)}|w|_{W^{-1,s}(\Omega)}.$$

If $w \in W^{1,s}(\Omega)$ there holds $\varphi = vw$. Moreover, $W^{1,s}(\Omega)$ is dense into $W^{-1,s}(\Omega)$. Therefore the product has a unique continuous expansion for $w \in W^{1,s}(\Omega)$, which is defined by $vw = \varphi$. □

Remark. A proof by duality gives weaker results than those of Lemma 3(ii). Particularly for $r = s = 2$ here we obtain $H^1(\Omega) \times H^{-1}(\Omega) \subset W^{-1,3/2}(\Omega)$. The duality only gives $H^1(\Omega) \times H^{-1}(\Omega) \subset W^{-1,3/2-\varepsilon}(\Omega)$, for all $\varepsilon > 0$.

Let us give some compactness properties for time-dependent functions. Denote by $\tau_h f$ the translated function of f, that is $(\tau_h f)(t) = f(t+h)$. Nikolskii spaces are defined for $1 \leqq q \leqq \infty$, $0 < s < 1$, by

$$N^{s,q}(0, T; E) = \{f \in L^q(0, T; E) : \sup_{h>0} h^{-s} \|\tau_h f - f\|_{L^q(0,T-h;E)} < \infty\}.$$

LEMMA 4. *Let* $X \subset E \subset Y$ *be Banach spaces, the imbedding* $X \to E$ *being compact. Then the following imbeddings are compact*:

(i) $$L^q(0, T; X) \cap \left\{\varphi : \frac{\partial \varphi}{\partial t} \in L^1(0, T; Y)\right\} \to L^q(0, T; E) \quad \text{if } 1 \leqq q \leqq \infty,$$

(ii) $$L^\infty(0, T; X) \cap \left\{\varphi : \frac{\partial \varphi}{\partial t} \in L^r(0, T; Y)\right\} \to C([0, T]; E) \quad \text{if } 1 < r \leqq \infty,$$

(iii) *For any given function* $k \in L^1(0, T)$, $k \geqq 0$ *and* $1 < r \leqq \infty$,

$$L^\infty(0, T; X) \cap \left\{\varphi : \left|\frac{\partial \varphi}{\partial t}\right|_Y - k \in L^r(0, T)\right\} \to C([0, T]; E),$$

(iv) $$L^q(0, T; X) \cap N^{s,q}(0, T; Y) \to L^q(0, T; E) \quad \text{if } s > 0, \quad 1 \leqq q \leqq \infty.$$

Proof. Parts (i) and (ii) are stated in Corollary 4 of [14, p. 85], and (iv) follows from Theorem 5 of [14, p. 84].

Proof of (iii). Let a family of functions φ satisfy

$$\left|\frac{\partial \varphi}{\partial t}\right|_Y \leqq k + \psi \quad \text{where } |\psi|_{L^r(0,T)} \leqq c.$$

Let $h>0$. By integration on $[0, t+h]$ it follows that

$$|\varphi(t+h)-\varphi(t)|_Y \leqq \int_t^{t+h} k(s)\,ds + h^{1-1/r}\left(\int_t^{t+h} |\psi(s)|^r\,ds\right)^{1/r}.$$

Taking the supremum for t in $[0, T-h]$, we obtain

$$|\tau_h\varphi-\varphi|_{L^\infty(0,\,T-h;\,Y)} \leqq \left(\sup_{0<t<T-h}\int_t^{t+h} k(s)\,ds\right) + ch^{1-1/r}.$$

The right-hand side goes to zero as $h\to 0$. Moreover, if the functions φ are bounded in $L^\infty(0, T; X)$ it follows, by Theorem 5 of [14, p. 84], that the set of functions φ is relatively compact in $C([0, T]; E)$. □

Let us give some estimates for ordinary differential equations.

LEMMA 5. *Let $g\in W^{1,1}(0, T)$, $g\geqq 0$ and $k\in L^1(0, T)$, $k\geqq 0$, satisfy*

$$\frac{d}{dt}g^2 \leqq gk, \qquad g(0)\leqq g_0.$$

Then

$$g(t)\leqq g_0+\frac{1}{2}\int_0^T k(s)\,ds \quad \forall t\in T.$$

Proof. Since $g\in W^{1,1}(0, T)$, g is continuous, $g(0)$ is defined, and $g_0\geqq 0$. Since $2g\,dg/dt\leqq gk$, there holds $2dg/dt\leqq k$ on the set $\{s: g(s)>0\}$.

Given t, let n be the larger time less than t such that $g(n)=g_0$. Then on $]n, t[$ there holds $2\,dg/dt\leqq k$; thus by integration on $[n, t]$ the announced inequality holds. If there is no such n, then $2\,dg/dt\leqq k$ on all of $[0, t]$ and the announced inequality follows by integration on $[0, t]$. □

LEMMA 6. *Let $g\in W^{1,1}(0, T)$ and $k\in L^1(0, T)$ satisfy*

$$\frac{dg}{dt}\leqq F(g)+k \quad \text{in } [0, T], \quad g(0)\leqq g_0$$

where F is bounded on bounded sets from $\mathbb{R}$ into $\mathbb{R}$, that is,

$$\forall a>0 \quad \exists A>0 \quad \text{such that } |x|\leqq a\Rightarrow |F(x)|\leqq A.$$

Then for every $\varepsilon>0$, there exists $T_\varepsilon>0$ independent of g such that

$$g(t)\leqq g_0+\varepsilon \quad \forall t\leqq T_\varepsilon.$$

Proof of Lemma 6. Since $g\in W^{1,1}(0, T)$, g is continuous. Let m be the smallest real value such that $g(m)=g_0+\varepsilon$ and let n be the largest real value less than m such that $g(n)=g_0$.

On $[n, m]$ there holds $F(g)\leqq A$ where $A=\sup\{F(x): g_0\leqq x\leqq g_0+\varepsilon\}$; then, by integrating the equation,

$$\varepsilon = g(m)-g(n)\leqq \int_n^m A+k(t)\,dt \leqq \int_0^m A+k^+(t)\,dt.$$

This proves that $m\geqq T_\varepsilon$, where T_ε is the smallest real value such that

$$\int_0^{T\varepsilon} A+k^+(t)\,dt=\varepsilon.$$

If this equation has no solution T_ε it proves that m does not exist, that is, $g\leqq g_0+\varepsilon$ on $(0, T)$, and thus the lemma is satisfied with $T_\varepsilon=T$. □

3. Weak t-smoothness of any solution. Equations (1)-(3) may be defined in the distribution sense once ρu and ρuu are defined, in particular, if $\rho \in L^\infty(Q)$, $u \in L^2(Q)$. In fact, we will obtain solutions that have the additional properties stated in (4).

Let us first prove that these properties (4) imply that ρ and $\int_\Omega \rho u \cdot v$ have integrable t-derivatives, and therefore are t-continuous. This continuity will be used for weak initial conditions to have a meaning.

PROPOSITION 7. *Assume*

$$\Omega \text{ is bounded and its boundary is Lipschitz,}$$

$$u \in L^2(0, T; V), \quad \rho \in L^\infty(Q), \quad p \in \mathcal{D}'(Q), \quad f \in L^1(0, T; (L^2(\Omega))^3), \tag{4}$$

$$\rho \geqq 0, \quad \rho^{1/2} u \in L^\infty(0, T; (L^2(\Omega))^3),$$

and let (1) *and* (2) *be satisfied.*

(i) *Then*

$$\frac{\partial \rho}{\partial t} \in L^2(0, T; W^{-1,6}(\Omega)) \cap L^\infty(0, T; H^{-1}(\Omega)),$$

$$\frac{\partial}{\partial t} \int_\Omega \rho u \cdot v \, dx \in L^1(0, T) \quad \forall v \in V.$$

More precisely, let $g = |\rho f|_{L^2} + |u\rho u - \mu \nabla u|_{L^2}$; *then* $g \in L^1(0, T)$ *and*

$$\left| \frac{\partial}{\partial t} \int_\Omega \rho u \cdot v \, dx \right| \leqq g |v|_V. \tag{5}$$

Therefore

$$\rho \in C([0, T]; W^{-1,\infty}(\Omega)),$$

$$\int_\Omega \rho u \cdot v \, dx \in C([0, T]) \quad \forall v \in V.$$

(ii) *In addition,*

$$\rho u \in L^2(0, T; (L^6(\Omega))^3) \cap L^\infty(0, T; (L^2(\Omega))^3),$$

$$\rho uu \in L^{4/3}(0, T; (L^2(\Omega))^9),$$

$$\frac{\partial \rho u}{\partial t} + \nabla \cdot (u\rho u) - \mu \Delta u - \rho f \in W^{-1,\infty}(0, T; (H^{-1}(\Omega))^3).$$

Proof of part (ii). By Sobolev's theorem $u \in L^2(0, T; (L^6(\Omega))^3)$. Moreover, $\rho \in L^\infty(Q) = L^\infty(0, T; L^\infty(\Omega))$; thus $\rho u \in L^2(0, T; (L^6(\Omega))^3)$.

There holds $\rho^{1/2} \in L^\infty(Q)$. Since $\rho^{1/2} u \in L^\infty(0, T; (L^2(\Omega))^3)$, thus $\rho u \in L^\infty(0, T; (L^2(\Omega))^3)$.

By Riesz's theorem, for every $0 \leqq s \leqq 1$, $\rho^{1/2} u \in L^a(0, T; (L^b(\Omega))^3)$ with $1/a = s/2 + (1-s)/\infty$, $1/b = s/6 + (1-s)/2$. Choosing $s = \frac{3}{4}$ we obtain $\rho^{1/2} u \in L^{8/3}(0, T; (L^4(\Omega))^3)$. Therefore $\rho uu \in L^{4/3}(0, T; (L^2(\Omega))^9)$.

Now $\partial \rho u / \partial t \in W^{-1,\infty}(0, T; (L^2(\Omega))^3)$. Moreover, $\nabla \cdot (u\rho u)$, $\mu \Delta u$, and ρf belong to $L^1(0, T; (H^{-1}(\Omega))^3)$, which is included in $W^{-1,\infty}(0, T; (H^{-1}(\Omega))^3)$ by Sobolev's theorem. This proves the last property of (ii). □

Proof of part (i) *of Proposition* 7. (a) By (ii) of Proposition 7 there holds

$$-\nabla \cdot (\rho u) \in L^2(0, T, W^{-1,6}(\Omega)) \cap L^\infty(0, T; H^{-1}(\Omega)).$$

By (2) this is $\partial \rho / \partial t$, which therefore belongs to these spaces.

(b) Then ρ is continuous on $[0, T]$ into $W^{-1,6}(\Omega)$. Since ρ is bounded into $L^\infty(\Omega)$, it is continuous into $L^\infty(\Omega)$ weak star, thus into $W^{-1,\infty}(\Omega)$.

(c) For every $v \in (\mathscr{D}(\Omega))^3$ equation (1) gives, in $\mathscr{D}'(]0, T[)$,

$$\left(\frac{\partial \rho u}{\partial t}, v\right)_\Omega = (\rho f - \nabla p - \nabla \cdot u\rho u + \mu \Delta u, v)_\Omega$$

$$= (\rho f, v)_\Omega + (p, \nabla \cdot v)_\Omega + (u\rho u - \mu \nabla u, \nabla v)_\Omega$$

and the left-hand term equals $(\partial/\partial t)(\rho u, v)_\Omega$.

For $v \in V$ it yields, since $\nabla \cdot v = 0$,

$$\frac{\partial}{\partial t} \int_\Omega \rho u \cdot v \, dx = \int_\Omega \rho f \cdot v + u\rho u \cdot \nabla v - \mu \nabla u \cdot \nabla v \, dx.$$

By continuity this equation holds for $v \in V$. Moreover, the right-hand side is bounded by $g|v|_V$, with

$$g = |\rho f|_{(L^2(\Omega))^3} + |u\rho u - \mu \nabla u|_{(L^2(\Omega))^9}. \tag{6}$$

By (ii), $g \in L^1(0, T)$. This proves (5). □

4. Fractional t-smoothness of any solution. In the preceding section we proved that $\int_\Omega \rho u \cdot v$ has an integrable t-derivative for every $v \in V$. This does not imply that ρu is t-smooth (for example, if $\rho u = \nabla q$, $q \in \mathscr{D}'(Q)$).

In this section we show that ρu has fractional t-smoothness, and more precisely that it lies in a Nikolskii space defined in § 1. This will be used later to obtain compactness in the existence proof.

PROPOSITION 8. (i) *Let the assumptions of Proposition 7 be satisfied. Then*

$$\rho u \in N^{1/4,2}(0, T; (W^{-1,3/2}(\Omega))^3).$$

(ii) *If in addition ρ has a lower bound $\alpha > 0$ in Q, then*

$$u \in N^{1/4,2}(0, T; (L^2(\Omega))^3).$$

Remark. In [13, p. 1012], Simon has obtained a similar result with $W^{-1,3/2-\varepsilon}(\Omega)$, for all $\varepsilon > 0$. Now we have $\varepsilon = 0$ since the properties of the product are improved by Lemma 3(iii).

Remark. By Proposition 8, ρu lies in a fractional Sobolev space:

$$\rho u \in W^{1/4-\varepsilon,2}(0, T; (W^{-1,3/2}(\Omega))^3) \quad \forall \varepsilon > 0.$$

Indeed for $0 < s < 1$ and $1 \leqq r \leqq \infty$ denote

$$W^{s,r}(0, T; E) = \left\{ f \in L^r(0, T; E) : \int_0^T \int_0^T \frac{|f(y) - f(x)|_E^r}{|y - x|^{1+sr}} \, dx \, dy < \infty \right\}.$$

Then, by Corollary 25 of [15], for all $\varepsilon > 0$,

$$N^{s,r}(0, T; E) \subset W^{s-\varepsilon,r}(0, T; E).$$

Remark. The continuity of ρu does not follow from Proposition 8. Indeed by the fractional Sobolev theorem for Nikolskii spaces (see, for example, [15, Cor. 28])

$$N^{s,r}(0, T; E) \subset C([0, T]; E) \quad \text{iff } s > 1/r.$$

Then the continuity would require $h^{1+2/\varepsilon}$ instead of $h^{1/4}$ in Proposition 8.

Proof of Proposition 8. In this proof c denotes various real numbers independent of h.

First step. For $h>0$, $v\in V$, and almost $t\in]0, T-h[$ there holds

$$\left(\int_\Omega \rho u\cdot v\,dx\right)(t+h)-\left(\int_\Omega \rho u\cdot v\,dx\right)(t)=\int_t^{t+h}\left(\frac{\partial}{\partial t}\int_\Omega \rho u\cdot v\,dx\right)(s)\,ds.$$

By (5) in Proposition 7, the right-hand side is bounded by

$$\left(\int_t^{t+h} g(s)\,ds\right)|v|_V.$$

Choosing $v=u(t+h)-u(t)$ and integrating in t, we find

$$\int_0^{T-h} dt\int_\Omega (\rho(t+h)u(t+h)-\rho(t)u(t))\cdot(u(t+h)-u(t))\,dx$$
$$\leqq \int_0^{T-h} dt\,|u(t+h)-u(t)|_V\int_t^{t+h} ds\,g(s).$$

By Fubini's theorem the right-hand term is equal to

$$=\int_0^T ds\,g(s)\int_{\overline{s-h}}^{\bar s} dt\,|u(t+h)-u(t)|_V$$

where $\bar s=0$ for $s\leqq 0$, $\bar s=s$ for $0\leqq s\leqq T-h$, $\bar s=T-h$ for $s\geqq T-h$. In this term let us bound

$$\int_{\overline{s-h}}^{\bar s}|u(t+h)-u(t)|_V\,dt\leqq\left(\int_{\overline{s-h}}^{\bar s}1^2\,ds\right)^{1/2}\left(\int_{\overline{s-h}}^{\bar s}|u(t+h)-u(t)|_V^2\,dt\right)^{1/2}$$
$$\leqq 2h^{1/2}|u|_{L^2(0,T;V)}.$$

Carrying back this estimate, we obtain

$$\int_0^{T-h} dt\int_\Omega (\rho(t+h)u(t+h)-\rho(t)u(t))\cdot(u(t+h)-u(t))\,dx\leqq ch^{1/2}. \tag{7}$$

Second step. For every $w\in\mathcal{D}(\Omega)$, (2) gives

$$\left(\frac{\partial\rho}{\partial t}, w\right)_\Omega=-(\nabla\cdot(\rho u), w)_\Omega=(\rho u, \nabla w)_\Omega,$$

which yields

$$\frac{\partial}{\partial t}\int_\Omega \rho w\,dx=\int_\Omega \rho u\cdot\nabla w\,dx.$$

Integrating we find for $h>0$ and for almost all t in $]0, T[$,

$$\int_\Omega(\rho(t+h)-\rho(t))w\,dx=\int_t^{t+h}\left(\int_\Omega(\rho u)(s)\cdot\nabla w\,dx\right)ds. \tag{8}$$

The right-hand side is bounded (using Hölder's inequality with $1=\frac16+\frac16+\frac23$) by

$$\leqq|\Omega|^{1/6}\left(\int_t^{t+h}|\rho(s)|_{L^\infty(\Omega)}|u(s)|_{(L^6(\Omega))^3}\,ds\right)|\nabla w|_{(L^{3/2}(\Omega))^3}$$
$$\leqq|\Omega|^{1/6}h^{1/2}\left(\int_t^{t+h}(|\rho(s)|_{L^\infty(\Omega)}|u(s)|_{(L^6(\Omega))^3})^2\,ds\right)^{1/2}|\nabla w|_{(L^{3/2}(\Omega))^3}$$
$$\leqq|\Omega|^{1/6}h^{1/2}c|\rho|_{L^\infty(Q)}|u|_{L^2(0,T;V)}|\nabla w|_{(L^{3/2}(\Omega))^3}.$$

By continuity this inequality is satisfied for $w \in W^{1,3/2}(\Omega)$. By Lemma 3(ii) we can choose $w = -u(t) \cdot (u(t+h) - u(t))$. Then

$$|\nabla w|_{(L^{3/2}(\Omega))^3} \leqq c|u(t)|_V |u(t+h) - u(t)|_V.$$

Carrying back these calculi into (8), and integrating in t, we find

$$\int_0^{T-h} dt \int_\Omega -(\rho(t+h) - \rho(t))u(t) \cdot (u(t+h) - u(t))\, dt \leqq ch^{1/2}. \tag{9}$$

Third step. Adding (7) and (9) we obtain

$$\int_0^{T-h} dt \int_\Omega \rho(t+h)(u(t+h) - u(t)) \cdot (u(t+h) - u(t))\, dx \leqq ch^{1/2}.$$

Since ρ is bounded in $L^\infty(Q)$, it follows that

$$\int_0^{T-h} dt \int_\Omega |\rho(t+h)(u(t+h) - u(t))|^2\, dx \leqq c|\rho|_{L^\infty(Q)} h^{1/2}.$$

Therefore

$$\|(\tau_h \rho)(\tau_h u - u)\|_{L^2(0,T-h;(L^2(\Omega))^3)} \leqq ch^{1/4}. \tag{10}$$

If $\rho \geqq \alpha$, then $\tau_h \rho \geqq \alpha$ and this gives

$$\|\tau_h u - u\|_{L^2(0,T-h;(L^2(\Omega))^3)} \leqq \left(\frac{c}{\alpha}\right) h^{1/4},$$

which proves (ii) of Proposition 8.

Fourth step. Equation (2) implies

$$\rho(t+h) - \rho(t) = -\int_t^{t+h} (\nabla \cdot (\rho u))(s)\, ds;$$

thus

$$\|\rho(t+h) - \rho(t)\|_{H^{-1}(\Omega)} \leqq h \sup_{0 \leqq s \leqq T} \|\nabla \cdot (\rho u)(s)\|_{H^{-1}(\Omega)}$$

and

$$\|\tau_h \rho - \rho\|_{L^\infty(0,T-h;H^{-1}(\Omega))} \leqq ch \|\rho u\|_{L^\infty(0,T;(L^2(\Omega))^3)}.$$

The product is continuous from $H^1(\Omega) \times H^{-1}(\Omega)$ into $W^{-1,3/2}(\Omega)$ by Lemma 3(iii). Therefore

$$\|(\tau_h \rho - \rho)u\|_{L^2(0,T-h;(W^{-1,3/2}(\Omega))^3)} \leqq c\|\tau_h \rho - \rho\|_{L^\infty(0,T-h;H^{-1}(\Omega))} \|u\|_{L^2(0,T;V)} \leqq ch.$$

Adding this inequality to (10), we finally obtain

$$\|\tau_h \rho \tau_h u - \rho u\|_{L^2(0,T-h;(W^{-1,3/2}(\Omega))^3)} \leqq ch^{1/4},$$

which proves (i) of Proposition 8. □

Remark. The first three steps are improvements by Simon [13] of estimates due to Antonzev and Kajikov [2]. The fourth step is an improvement of [13].

5. Existence of a solution satisfying the weak initial condition. In this section we give an existence result for a solution of the model stated in § 1, except that the exact initial condition $(\rho u)(0) = \rho_0 u_0$ is replaced by a weaker one. Sufficient conditions that guarantee the exact initial condition is satisfied will be given in §§ 6–8.

THEOREM 9. *Let the following hold*:

$$\Omega \text{ is bounded and its boundary } \Gamma \text{ is Lipschitz,}$$

$$f \in L^1(0, T; (L^2(\Omega))^3), \quad u_0 \in H, \quad \rho_0 \in L^\infty(\Omega), \quad \rho_0 \geqq 0.$$

(i) *There exist*

$$u \in L^2(0, T; V), \quad \rho \in L^\infty(Q), \quad p \in W^{-1,\infty}(0, T; L^2(\Omega))$$

such that

$$\inf_\Omega \rho_0 \leqq \rho \leqq \sup_\Omega \rho_0,$$

$$\rho u \in L^\infty(0, T; (L^2(\Omega))^3) \cap N^{1/4,2}(0, T; (W^{-1,3/2}(\Omega))^3)$$

$$\rho \in C([0, T]; W^{-1,\infty}(\Omega)), \quad \int_\Omega \rho u \cdot v \, dx \in C([0, T]) \quad \forall v \in V,$$

which satisfy (1)–(3) *and the initial conditions*

$$\rho(0) = \rho_0,$$

$$\left(\int_\Omega \rho u \cdot v \, dx\right)(0) = \int_\Omega \rho_0 u_0 \cdot v \, dx \quad \forall v \in V. \tag{11}$$

Moreover, u, ρ, p satisfy all the properties stated in Propositions 7 and 8.

(ii) *If in addition*

$$\inf_\Omega \rho_0 > 0,$$

then

$$u \in L^\infty(0, T; (L^2(\Omega))^3) \cap N^{1/4,2}(0, T; (L^2(\Omega))^3).$$

Remark. The boundary condition $u = 0$ on $\Gamma \times]0, T[$ in the trace sense is implied by $u \in L^2(0, T; V)$.

Remark. Variational equation. Antonzev and Kajikov [2], Lions [10], or Simon [13] solve, instead of (1) and of the weak initial condition (11), the following variational equation:

$$\int_Q \left(-\rho u \frac{\partial \varphi}{\partial t} - u\rho u \cdot \nabla \varphi + \mu \nabla u \cdot \nabla \varphi - \rho f \cdot \varphi\right) dx \, dt = \int_\Omega \rho_0 u_0 \cdot \varphi(0) \, dx \tag{12}$$

$$\forall \varphi \in C^1([0, T]; V) \text{ such that } \varphi(T) = 0.$$

By integration by parts, we can prove that the solution of Theorem 9 satisfies (12). Conversely, from (12), we can obtain the pressure and therefore (1) by using Lemma 2, and we can obtain (11) by integration by parts.

Therefore this variational equation is, in fact, equivalent to (1) and to the initial condition (11).

Remark. Antonzev and Kajikov [2], Lions [10], and Simon [13] suppose that $f \in L^2(0, T; H)$. In fact their proofs hold for $f \in L^2(0, T; (L^2(\Omega))^3)$; they do not require f to be divergence free. Here we conclude with only $f \in L^1(0, T; (L^2(\Omega))^3)$.

Indeed to estimate the approached solutions u^m, ρ^m from (18) (which is given in the following proof) they use Gronwall's lemma. Here we improve the estimates by using Lemma 5.

Remark. The sketch of the following proof is mainly due to Antonzev and Kajikov [2]. The obtainment of the pressure and of the weak initial condition on ρu are new, and several estimates are improved.

Remark. An uniqueness result of a solution u, ρ of the exact model in a class of more regular functions is given by Ladyzhenskaya and Solonnikov [18]. An existence result in an unbounded domain Ω is given by Padula [19].

Proof of Theorem 9. *Definition of approached solutions* u^m, ρ^m. Let $w^1, \cdots, w^m, \cdots$ be a basis of the Hilbert space V such that

$$w^m \in (C^1(\bar{\Omega}))^3. \tag{13}$$

Such a basis exists since $(C^1(\bar{\Omega}))^3 \cap V$ is dense in V. Let

$$V^m = \text{subspace of } V \text{ generated by } (w^1, \cdots, w^m).$$

We are looking for u^m, ρ^m such that, for some $T_m > 0$,

$$u^m \in C^1([0, T_m], V^m), \qquad \rho^m \in C^1([0, T_m], C^1(\bar{\Omega})),$$

$$\int_\Omega \left(\left(\frac{\partial \rho^m u^m}{\partial t} + \nabla \cdot (u^m \rho^m u^m) - \rho^m f \right) \cdot v + \mu \nabla u^m \cdot \nabla v \right) dx = 0 \quad \forall v \in V^m, \tag{14}$$

$$\frac{\partial \rho^m}{\partial t} + \nabla \cdot (\rho^m u^m) = 0, \tag{15}$$

$$u^m(0) = u_0^m, \qquad \rho^m(0) = \rho_0^m$$

where u_0^m and ρ_0^m are any functions satisfying

$$\begin{aligned} &u_0^m \in V^m, \quad u_0^m \to u_0 \quad \text{in } (L^2(\Omega))^3, \\ &\rho_0^m \in C^1(\bar{\Omega}), \quad \rho_0^m \to \rho_0 \quad \text{in } L^\infty(\Omega) \text{ weak star as } m \to \infty, \\ &\frac{1}{m} + \inf_\Omega \rho_0 \leqq \rho_0^m \leqq \frac{1}{m} + \sup_\Omega \rho_0. \end{aligned} \tag{16}$$

Equations (14), (15) may be replaced by the nonconservative ones:

$$\int_\Omega \left(\rho^m \left(\frac{\partial u^m}{\partial t} + (u^m \cdot \nabla) u^m - f \right) \cdot v + \mu \nabla u^m \cdot \nabla v \right) dx = 0 \quad \forall v \in V^m, \tag{14'}$$

$$\frac{\partial \rho^m}{\partial t} + u^m \cdot \nabla \rho^m = 0. \tag{15'}$$

Indeed (15') and (15) differ from ρ^m times $\nabla \cdot u^m$, which is zero, and (14') and (14) differ from the integral over Ω of $\cdot\, v$ times (15).

Local existence of u^m, ρ^m. Assuming that u^m exists, the trajectory $y^m = y^m_{x,t}(s)$ of a particle located in x at time t is defined by

$$\frac{dy^m}{ds}(s) = u^m(y^m(s), s) \quad \forall s \geqq 0, \quad y^m(t) = x.$$

By (13) there holds $u^m \in C^1([0, T]; (C^1(\bar{\Omega}))^3)$; thus y^m lies in the same space, and the map $\Lambda : u^m \to y^m$ is continuous.

For a fixed u^m, (15') has a unique solution ρ^m such that $\rho^m(0) = \rho_0^m$, which is

$$\rho^m(x, t) = \rho_0^m(y^m_{x,t}(t)).$$

This yields

$$\rho^m = \rho_0^m(\Lambda u^m).$$

In (14′) let us use this and decompose

$$u^m = \varphi_1^m w^1 + \cdots + \varphi_m^m w^m, \qquad \varphi_j^m \in C^1([0, T_m]).$$

Then choosing successively $v = w^1, w^2, \cdots, w^m$ we obtain the following system of m nonlinear ordinary differential equations on $\varphi_j^m, j = 1, \cdots, m$:

$$F_j^m(\varphi_1^m, \cdots, \varphi_m^m)\frac{\partial \varphi_j^m}{\partial t} + G_j^m(\varphi_1^m, \cdots, \varphi_m^m) = 0.$$

The functions F_j^m and G_j^m are continuous, and

$$F_j^m(\varphi_1^m, \cdots, \varphi_m^m)(t) = \int_\Omega \rho_0^m((\Lambda u^m)(x, t))\, dx \geqq \frac{\text{meas. } \Omega}{m}$$

since $\rho_0^m \geqq 1/m$.

Therefore this system is equivalent to

$$\frac{\partial \varphi_j^m}{\partial t} = -\left(\frac{G_j^m}{F_j^m}\right)(\varphi_1^m, \cdots, \varphi_m^m), \qquad j = 1, \cdots, m$$

where G_j^m/F_j^m is continuous from $(C^1([0, T_m]))^m$ into $C^1([0, T_m])$.

Such a system has a local solution $\varphi_1^m, \cdots, \varphi_m^m$, that is, a solution for some $T_m > 0$, such that

$$\varphi_1^m(0) = \varphi_{j0}^m \quad \text{where } u_0^m = \varphi_{10}^m w^1 + \cdots + \varphi_{m0}^m w^m.$$

The corresponding u^m, ρ^m satisfy (14′), (15′) and therefore (14), (15), and the initial conditions.

Global existence and estimates on u^m, ρ^m. The above expression of ρ^m and the choice of ρ_0^m yield, on $[0, T_m]$,

$$\frac{1}{m} + \inf_\Omega \rho_0 \leqq \rho^m \leqq \frac{1}{m} + \sup_\Omega \rho_0. \tag{17}$$

At any time t multiply (15) by $-|u^m(t)|^2$ and integrate over Ω. And add to (14) with $v = 2u^m(t)$. This gives

$$\int_\Omega \left(\frac{d}{dt}(\rho^m|u^m|^2) + \nabla \cdot (\rho^m u^m |u^m|^2) - 2\rho^m f \cdot u^m + 2\mu|\nabla u^m|^2\right) dx = 0.$$

The integral of the second term equals zero since $u^m = 0$ on the boundary Γ. Bounding the third term by Hölder's inequality and by $\rho^m \leqq b = 1 + \sup_\Omega \rho_0$, we find

$$\frac{d}{dt}\int_\Omega \rho^m|u^m|^2\, dx + 2\mu \int_\Omega |\nabla u^m|^2\, dx \leqq 2b^{1/2}\left(\int_\Omega \rho^m|u^m|^2\, dx\right)^{1/2}\left(\int_\Omega |f|^2\, dx\right)^{1/2}. \tag{18}$$

By the choice of ρ_0^m and u_0^m there exists d independent of m such that

$$\left(\int_\Omega \rho^m|u^m|^2\, dx\right)(0) = \int_\Omega \rho_0^m|u_0^m|^2\, dx \leqq d.$$

It follows by Lemma 5 that, for all $t \leqq T_m$,

$$\left(\int_\Omega \rho^m|u^m|^2\, dx\right)(t) = d + b^{1/2}\int_0^T \left(\int_\Omega |f(x, s)|^2\, dx\right)^{1/2} ds.$$

It follows that $T_m = T$ and that

$$(\rho^m)^{1/2} u^m \text{ is bounded in } L^\infty(0, T; (L^2(\Omega))^3).$$

By integrating (18) on $[0, T]$, and by (17),

$$u^m \text{ is bounded in } L^2(0, T; V),$$

$$\rho^m \text{ is bounded in } L^\infty(0, T; L^\infty(\Omega)),$$

and thus

$$\rho^m u^m \text{ is bounded in } L^\infty(0, T; (L^2(\Omega))^3).$$

In addition all the properties proved in §§ 3 and 4 for the exact solutions u, ρ are satisfied by the approached solutions u^m, ρ^m with norms independent of m. This is obtained by replacing u, ρ by u^m, ρ^m in the proofs of Propositions 7 and 8. In particular,

$$u^m \rho^m u^m \text{ is bounded in } L^{4/3}(0, T; (L^2(\Omega))^9),$$

$$\frac{\partial \rho^m}{\partial t} \text{ is bounded in } L^2(0, T; W^{-1,6}(\Omega)),$$

$$\left| \frac{\partial}{\partial t} \int_\Omega \rho^m u^m \cdot v \, dx \right| \leqq (|\rho^m f|_{(L^2(\Omega))^3} + |u^m \rho^m u^m - \mu \nabla u^m|_{(L^2(\Omega))^9}) |v|_V,$$

$$\rho^m u^m \text{ is bounded in } N^{1/4,2}(0, T; (W^{-1,3/2}(\Omega))^3).$$

Convergence properties. Using Lemma 4(ii) with $X = L^\infty(\Omega)$, $E = W^{-1,\infty}(\Omega)$, $Y = W^{-1,6}(\Omega)$, and $r = 2$, the estimates on ρ^m imply that

$$\{\rho^m\}_{m \in \mathbb{N}} \text{ is relatively compact in } C([0, T]; W^{-1,\infty}(\Omega)).$$

Using Lemma 4 (iv) with $X = (L^2(\Omega))^3$, $E = (H^{-1}(\Omega))^3$, $Y = (W^{-1,3/2}(\Omega))^3$, $s = \frac{1}{4}$, and $q = 2$, the estimates on $\rho^m u^m$ imply that

$$\{\rho^m u^m\}_{m \in \mathbb{N}} \text{ is relatively compact in } L^2(0, T; (H^{-1}(\Omega))^3).$$

Therefore there exists a subsequence of $\{u^m, \rho^m\}_{m \in \mathbb{N}}$ and u, ρ, g, k such that

$u^m \to u$ in $L^2(0, T; V)$ weak,

$\rho^m \to \rho$ in $L^\infty(Q)$ weak star and in $C([0, T]; W^{-1,\infty}(\Omega))$ strong,

$\rho^m u^m \to g$ in $L^\infty(0, T; (L^2(\Omega))^3)$ weak star and in $L^2(0, T; (H^{-1}(\Omega))^3)$ strong,

$\rho^m u^m u^m \to k$ in $L^{4/3}(0, T; (L^2(\Omega))^9)$ weak.

By Lemma 3 (iii) the product is continuous from $H_0^1 \times W^{-1,\infty}$ into $W^{-1,6}$. Therefore these first two properties imply that

$$\rho^m u^m \to \rho u \text{ in } L^2(0, T; (W^{-1,6}(\Omega))^3) \text{ weak},$$

and thus $g = \rho u$ in the third property.

From the first and the third properties it follows, again using Lemma 3(iii), that

$$\rho^m u^m u^m \to g u \text{ in } L^1(0, T; (W^{-1,3/2}(\Omega))^9) \text{ weak},$$

and thus $k = \rho u u$ in the fourth property.

Limit equations and initial conditions.

(i) There holds $\rho^m \to \rho$ in $\mathcal{D}'(Q)$ and $\rho^m u^m \to \rho u$ in $(\mathcal{D}'(Q))^3$. Then passing to the limit in $\mathcal{D}'(Q)$ in (15), we find

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho u) = 0.$$

(ii) There holds $\rho^m(0) \to \rho(0)$ in $W^{-1,\infty}(\Omega)$. Then passing to the limit in the initial condition $\rho^m(0) = \rho_0^m$ we obtain

$$\rho(0) = \rho_0.$$

(iii) Since $u \in L^2(0, T; V)$, the following equation holds:

$$\nabla \cdot u = 0.$$

(iv) Equation (14) may be written, for all $w \in V^m$,

$$\frac{\partial}{\partial t}\int_\Omega \rho^m u^m \cdot w\, dx + \int_\Omega (-u^m \rho^m u^m + \mu \nabla u^m) \cdot \nabla w - \rho^m f \cdot w\, dx = 0.$$

Let $v \in V$ and choose $w = v^m$ where $v^m \to v$ in V. Thus we can pass to the limit in each term in $\mathscr{D}'(]0, T[)$. It gives

$$\frac{\partial}{\partial t}\int_\Omega \rho u \cdot v\, dx + \int_\Omega (-u\rho u + \mu \nabla u) \cdot \nabla v - \rho f \cdot v\, dx = 0.$$

This yields

$$\left(\frac{\partial \rho u}{\partial t} + \nabla \cdot (u \rho u) - \mu \Delta u - \rho f, v\right)_\Omega = 0 \quad \forall v \in V.$$

Moreover by Proposition 7(ii) $\partial \rho u/\partial t + \nabla \cdot (u\rho u) - \mu \Delta u - \rho f \in W^{-1,\infty}(0, T; (H^{-1}(\Omega))^3)$. Then by Lemma 2 there exists $p \in W^{-1,\infty}(0, T; L^2(\Omega))$ such that

$$\frac{\partial \rho u}{\partial t} + \nabla \cdot (u \rho u) - \mu \Delta u - \rho f = \nabla p.$$

(v) *Let* $v \in V$ and $v^m \in V^m$ satisfy $v^m \to v$ in V. By the above estimates

$$\left|\frac{\partial}{\partial t}\int_\Omega \rho^m u^m \cdot v^m\, dx\right| \leqq k + \psi^m,$$

$$k = b \sup_m |v^m|_V, \qquad \psi^m = |u^m \rho^m u^m - \mu \nabla u^m|_{(L^2(\Omega))^9} |v^m|_V.$$

There holds $k \in L^1(0, T)$ and ψ^m is bounded in $L^{4/3}(0, T)$. Therefore Lemma 4(iii) with $X = E = Y = \mathbb{R}$ and $r = 4/3$ gives

$$\int_\Omega \rho^m u^m \cdot v^m\, dx \text{ is relatively compact in } C([0, T]).$$

On the other hand, this sequence converges to $\int_\Omega \rho u \cdot v\, dx$ in $L^2(0, T)$ weak, by the convergence properties. Thus the convergence holds in $C([0, T])$ and

$$\left(\int_\Omega \rho^m u^m \cdot v^m\, dx\right)(0) \to \left(\int_\Omega \rho u \cdot v\, dx\right)(0).$$

The left-hand term is

$$\int_\Omega \rho_0^m u_0^m \cdot v^m\, dx \to \int_\Omega \rho_0 u_0 \cdot v\, dx.$$

Thus

$$\left(\int_\Omega \rho u \cdot v\, dx\right)(0) = \int_\Omega \rho_0 u_0 \cdot v\, dx \quad \forall v \in V.$$

(vi) Passing to the limit in (17), we obtain

$$\inf_{\Omega} \rho_0 \leqq \rho \leqq \sup_{\Omega} \rho_0.$$

(vii) Proposition 8(i) gives

$$\rho u \in N^{1/4,2}(0, T; (W^{-1,3/2}(\Omega))^3),$$

which ends the proof of (i) of Theorem 9.

(viii) If $\inf_\Omega \rho_0 > 0$, since by (vi) of this proof $\rho \geqq \inf_\Omega \rho_0$, then $1/\rho \in L^\infty(Q)$. By (i) of Theorem 9, $\rho u \in L^\infty(0, T; (L^2(\Omega))^3)$. Therefore $u = \rho u \times 1/\rho$ satisfies

$$u \in L^\infty(0, T; (L^2(\Omega))^3).$$

Moreover, by Proposition 8(ii)

$$u \in N^{1/4,2}(0, T; (L^2(\Omega))^3),$$

which ends the proof of part (ii) of Theorem 9. □

6. Navier–Stokes equations. The Navier-Stokes equations are obtained by setting $\rho \equiv 1$. In this case we will prove that the strong initial condition $u(0) = u_0$ is satisfied. Indeed the t-continuity of u in a certain space V_* will result from the estimate on $(\partial/\partial t) \int_\Omega \rho u \cdot v \, dx$ proved in § 3.

For every $v \in (L^2(\Omega))^3$ let

$$|v|_* = \sup_{\varphi \in V, \varphi \neq 0} \frac{1}{|\varphi|_V} \left| \int_\Omega v \cdot \varphi \, dx \right|. \tag{19}$$

LEMMA 10. *$|\ |_*$ is a norm on V, and is only a seminorm on $(H^1(\Omega))^3$.*

Proof. If $v \in V$ and $|v|_* = 0$, then choosing $\varphi = v$ we obtain $v = 0$. Thus $|\ |_*$ is a norm on V.

Let $v = \nabla w$, where $w \in (\mathscr{D}(\Omega))^3$, $w \neq 0$. Then

$$\int_\Omega v \cdot \varphi \, dx = -\int_\Omega w \nabla \cdot \varphi \, dx = 0 \quad \forall \varphi \in V,$$

and $|v|_* = 0$. Thus $|\ |_*$ is not a norm on $(H^1(\Omega))^3$. □

By definition V_* is the completion of V for the norm $|\ |_*$.

PROPOSITION 11. *Let*

$$f \in L^1(0, T; (L^2(\Omega))^3), \qquad u_0 \in H.$$

There exists u, p such that

$$u \in L^2(0, T; V) \cap L^\infty(0, T; H) \cap C([0, T]; (H^{-1}(\Omega))^3),$$

$$\frac{\partial u}{\partial t} \in L^1(0, T; V_*), \qquad p \in W^{-1,\infty}(0, T; L^2(\Omega)),$$

which satisfy the Navier–Stokes equations, that is, (1) *with $\rho \equiv 1$ and* (2), *and the initial condition*

$$u(0) = u_0.$$

This result is given to point out the difference between the homogeneous and nonhomogeneous cases. For only the homogeneous case the assumption of f may be replaced by $f \in L^2(0, T; (H^{-1}(\Omega))^3)$ (see, for example, [16]).

Remark. The space V_* may be replaced by the dual space V' of V. Indeed when H is identified to its dual space H', then V_* is identified to V'.

But when these identifications are done, it is not possible to identify at the same time H to a subspace of $(L^2(\Omega))^3$ and of $(H^{-1}(\Omega))^3$.

Remark. The space V_* contains some v that are not in $(\mathscr{D}'(\Omega))^3$. For a sequence v_n of V the convergence in $(H^{-1}(\Omega))^3$ implies the convergence for $|\ |_*$, but the convergence for $|\ |_*$ does not imply the convergence in $(\mathscr{D}'(\Omega))^3$. These properties are proved in Simon [16].

Proof of Proposition 11. (i) Choosing $\rho_0 \equiv 1$ the proof of Theorem 9 is satisfied with $\rho^m \equiv 1$. Therefore Theorem 9 is satisfied with $\rho \equiv 1$. Then the property (5) in Proposition 7 yields

$$\left|\frac{\partial}{\partial t}\int_\Omega u \cdot v\, dx\right| \leqq g|v|_V \quad \text{with } g \in L^1(0, T). \tag{20}$$

Since $u \in L^2(0, T; V)$ we have $\partial u/\partial t \in \mathscr{D}'(0, T; V)$ and this inequality yields $|\partial u/\partial t|_{V_*} \leqq g$. Thus

$$\frac{\partial u}{\partial t} \in L^1(0, T, V_*).$$

(ii) Then u is continuous on $[0, T]$ into V_*. Since u is bounded into H, it is continuous into $(L^2(\Omega))^3$ equipped with the weak topology and therefore into $(H^{-1}(\Omega))^3$ strong.

(iii) It follows that $u(0) \in H$, and that

$$\left(\int_\Omega u \cdot v\, dx\right)(0) = \int_\Omega u(0) \cdot v\, dx.$$

Then the weak initial condition (11) yields

$$\left(\int_\Omega (u(0) - u_0) \cdot v\, dx\right) = 0 \quad \forall v \in V.$$

By continuity it is satisfied for all $v \in H$, thus for $v = u(0) - u_0$, which gives

$$u(0) - u_0 = 0. \qquad \square$$

7. Sufficient conditions for the strong initial condition. Using the seminorm $|\ |_*$ defined in § 6, the estimate (5) proved in Proposition 7 yields

$$\left|\frac{\partial \rho u}{\partial t}\right|_* \leqq g \quad \text{where } g \in L^1(0, T). \tag{21}$$

If $\rho u \in \mathscr{D}'(0, T; V_*)$ this gives the t-continuity of ρu into V^*, $|\ |_*$ being a norm on V_*. For the Navier-Stokes equations, this is satisfied since $\rho u = u$.

Unfortunately in the nonhomogeneous case the following holds.

LEMMA 12. *If ρ depends on t, then $\rho u \notin \mathscr{D}'(0, T; V_*)$ and* (21) *is not enough to define* $(\rho u)(0)$.

Proof of Lemma 12. By continuity every $v \in V_*$ satisfies $\nabla \cdot v = 0$. If $\rho u \in \mathscr{D}'(0, T; V_*)$, then $\nabla \cdot (\rho u) = 0$ and (2) gives $\partial \rho/\partial t = 0$, so that ρ is independent of t.

Since ρu is only known to be in $L^\infty(0, T; (L^2(\Omega))^3)$, and $|\ |_*$ is not a norm on $(L^2(\Omega))^3$, (21) does not allow us to define $(\rho u)(0)$. Remark that, if a seminorm were enough, then $\varphi(0)$ would be defined for every function by using the null seminorm! $\square$

Now we are looking for the strong initial conditions that are satisfied when ρu or u are continuous.

PROPOSITION 13. *Let u, ρ, p be a solution given by Theorem* 9(i).

(i) *If $\rho u \in C([0, T]; X)$, where X is any Banach space, then $(\rho u)(0) \in (L^2(\Omega))^3$ and*

$$\int_\Omega ((\rho u)(0) - \rho_0 u_0) \cdot v \, dx \quad \forall v \in H.$$

Therefore

$$(\rho u)(0) = \rho_0 u_0 + \nabla \lambda \quad \text{where } \lambda \in H^1(\Omega).$$

(ii) *If $u \in C([0, T]; H)$, then*

$$(\rho u)(0) = \rho(0) u(0) = \rho_0 u_0.$$

(iii) *If $u \in C([0, T]; H)$ and $\rho_0 > 0$ almost everywhere in Ω, then*

$$u(0) = u_0.$$

Proof. (i) If ρu is continuous on $[0, T]$ into X, it is continuous into $(L^2(\Omega))^3$ weak since it is bounded into $(L^2(\Omega))^3$ by Theorem 9(i). Then $(\rho u)(0) \in (L^2(\Omega))^3$,

$$\left(\int_\Omega \rho u \cdot v \, dx \right)(0) = \int_\Omega (\rho u)(0) \cdot v \, dx$$

and the weak initial condition (11) yields

$$\int_\Omega ((\rho u)(0) - \rho_0 u_0) \cdot v \, dx = 0 \quad \forall v \in V.$$

By continuity this is satisfied by all $v \in H$.

By Lemma 1 there exists $\lambda \in L^2(\Omega)$ such that $(\rho u)(0) = \rho_0 u_0 + \nabla \lambda$. Then $\lambda \in H^1(\Omega)$.

(ii) Let u be continuous on $[0, T]$ into H. By Theorem 9(i), ρ is continuous into $W^{-1,\infty}(\Omega)$ and is bounded into $L^\infty(\Omega)$, thus it is continuous into $L^\infty(\Omega)$ weak star. Thus ρu is continuous into $(L^2(\Omega))^3$ weak, and $(\rho u)(0) = \rho(0) u(0)$. Then, by (i) of this proposition,

$$\int_\Omega (\rho(0) u(0) - \rho_0 u_0) \cdot v \, dx = 0 \quad \forall v \in H.$$

Choosing $v = u(0) - u_0$ and using $\rho(0) = \rho_0$ it gives

$$\rho_0(x) |u(0) - u_0|^2(x) = 0 \quad \text{for almost all } x \in \Omega. \tag{22}$$

Multiplying by ρ_0 we obtain $|\rho_0 (u(0) - u_0)|^2 = 0$, and thus

$$\rho(0) u(0) - \rho_0 u_0 = 0.$$

(iii) If moreover $\rho_0(x) > 0$ for almost all $x \in \Omega$, then (22) gives

$$u(0) - u_0 = 0 \quad \text{almost everywhere in } \Omega. \qquad \square$$

8. Existence of a solution satisfying the strong initial condition. In this section we will prove that, when u_0, f, and Ω are regular enough, there exists a global solution that is regular up to some time T_*. Then ρu satisfies an initial condition that is intermediate between the strong and the weak ones. When in addition ρ_0 is not too small, the strong initial condition is satisfied.

THEOREM 14. *Assume*

$$\Omega \text{ is bounded and its boundary is } W^{2,\infty} \text{ (or } C^2),$$
$$f \in (L^2(Q))^3, \quad u_0 \in V, \quad \rho_0 \in L^\infty(\Omega), \quad \rho_0 \geqq 0.$$

(i) *There exist* u, ρ, p *that satisfy all the properties stated in Theorem* 9 *and, for some* $T_* > 0$,

$$u \in L^2(0, T_*; (H^2(\Omega))^3) \cap L^\infty(0, T_*; V), \qquad p \in L^2(0, T_*; L^6(\Omega)),$$

$$\rho u \in C([0, T_*]; (W^{-1,\infty}(\Omega))^3), \qquad \frac{\partial \rho u}{\partial t} \in L^2(0, T_*; (W^{-1,6}(\Omega))^3),$$

$$\frac{\partial \rho}{\partial t} \in L^\infty(0, T_*; W^{-1,6}(\Omega)) \cap L^{4-\varepsilon}(0, T_*; W^{-1,\infty}(\Omega)) \quad \forall \varepsilon > 0,$$

and the initial condition

$$(\rho u)(0) = \rho_0 u_0 + \nabla \lambda \quad \textit{where } \lambda \in W^{1,6}(\Omega).$$

The time T_* *depends only on the data* Ω, f, u_0, ρ_0.

(ii) *If in addition*

$$\frac{1}{\rho_0} \in L^{6/5}(\Omega),$$

then

$$t \to u(t) \textit{ is continuous at } t = 0 \textit{ into } H,$$

$$u(0) = u_0, \qquad (\rho u)(0) = \rho_0 u_0.$$

(iii) *If in addition*

$$\inf_\Omega \rho_0 > 0,$$

then

$$u \in C([0, T_*]; V), \qquad \frac{\partial u}{\partial t} \in L^2(0, T_*; H).$$

Remark. The assumption $1/\rho_0 \in L^{6/5}(\Omega)$ allows ρ_0 to have some zeros, but the set of all zeros must have a null measure. This assumption is particularly satisfied if $\inf_\Omega \rho_0 > 0$.

Let P denote the projection in $(L^2(\Omega))^3$ on H. We remark that

$$(P\Delta v, w)_\Omega = -(\nabla v, \nabla w)_\Omega \quad \forall v \in (H^2(\Omega))^3 \quad \forall w \in V.$$

Since the boundary Γ is $W^{2,\infty}$, the following estimates hold.

LEMMA 15. *There exist* $e > 0$ *and* $c > 0$ *such that, for all* $v \in (H^2(\Omega))^3 \cap V$,

(i) $|\Delta v|_{(L^2(\Omega))^3} \leqq e |P\Delta v|_{(L^2(\Omega))^3}$,

(ii) $|v|_{(L^\infty(\Omega))^3} \leqq c(|\Delta v|_{(L^2(\Omega))^3})^{3/4}(|\nabla v|_{(L^2(\Omega))^9})^{1/4}$.

Proof. (i) There holds $(\Delta v - P\Delta v, w)_\Omega = 0$, for all $w \in H$. Thus by Lemma 1 there exists $q \in L^2(\Omega)$ such that $\Delta v = P\Delta v + \nabla q$. Inequality (i) is then given by a theorem of Cattabriga [3] (see equally [8, Thm. 2, p. 65]).

(ii) By Riesz's theorem with $\frac{1}{4} = s/6 + (1-s)/2$, then $s = \frac{3}{4}$, there holds

$$|v|_{(W^{1,4}(\Omega))^3} \leqq c(|v|_{(W^{1,6}(\Omega))^3})^{3/4}(|v|_{(W^{1,2}(\Omega))^3})^{1/4}.$$

By Sobolev's theorem $W^{1,4}(\Omega) \subset L^\infty(\Omega)$ and $H^2(\Omega) \subset W^{1,6}(\Omega)$. Inequality (ii) follows, since $|v|_{H^2} \leqq c'|\Delta v|_{L^2}$. □

Proof of part (i) *of Theorem* 14.

Special basis. We use a basis of eigenfunctions still used for Navier-Stokes equations by Lions [11, p. 74] and by Ladyzhenskaya [8, pp. 43-45]. The operator $P\Delta$ is a bijection of $(H^2(\Omega))^3 \cap V$ on H that is self-adjoint and whose inverse is compact [8, Thms. 6 and 2, pp. 44, 65]. Thus there exists an increasing sequence of positive eigenvalues λ_m and a sequence of corresponding eigenfunctions w_m defined by

$$w_m \in (H^2(\Omega))^3 \cap V, \qquad P\Delta w_m + \lambda_m w_m = 0.$$

The set $\{w_m\}_{m\in N}$ is an orthogonal basis in H and a basis in V [8, p. 45]. By Sobolev's theorem there holds $P\Delta w_m = \lambda_m w_m \in (C(\bar{\Omega}))^3$. Therefore [8, Thm. 2, p. 74] $w_m \in (W^{2,q}(\Omega))^3$, for all $q<\infty$, which implies

$$w_m \in (C^1(\bar{\Omega}))^N.$$

This is assumption (13), and thus we can use this basis in the proof of Theorem 9. Then the approached solutions u^m, ρ^m satisfy all the estimates obtained in the proof of Theorem 9.

Supplementary estimate. In (14′) for any t it is possible to choose $v = 2(\partial u^m/\partial t)(t)$. Indeed it belongs to V^m. Then (14′) yields

$$2\int_\Omega \rho^m \left|\frac{\partial u^m}{\partial t}\right|^2 dx + \mu \frac{d}{dt}\int_\Omega |\nabla u^m|^2 = 2\int_\Omega \rho^m(-(u^m\cdot\nabla)u^m + f)\cdot\frac{\partial u^m}{\partial t}\,dx$$
$$\leqq \int_\Omega \rho^m \left|\frac{\partial u^m}{\partial t}\right|^2 dx + \int_\Omega \rho^m |(u^m\cdot\nabla)u^m - f|^2\,dx.$$

On the other hand, for any t it is possible to choose $v = P\Delta u^m(t)$ in (14′). Indeed it belongs to V^m since $P\Delta$ maps V^m into V^m by $P\Delta w_m = -\lambda_m w_m$. Then (14′) yields

$$\mu\int_\Omega \Delta u^M\cdot P\Delta u^m\,dx = \int_\Omega \rho^m\left(\frac{\partial u^m}{\partial t} + (u^m\cdot\nabla)u^m - f\right)\cdot P\Delta u^m\,dx$$
$$\leqq \frac{\mu}{2}\int_\Omega |P\Delta u^m|^2\,dx$$
$$+\frac{1}{2\mu}\int_\Omega 2|\rho^m|^2\left(\left|\frac{\partial u^m}{\partial t}\right|^2 + |(u^m\cdot\nabla)u^m - f|^2\right)dx.$$

Let us multiply this inequality by d and add it to the preceding inequality. Then observe that, by definition of P, the integral on Ω of $\Delta u^m\cdot P\Delta u^m$ equals the integral of $|P\Delta u^m|^2$, and that, by (17), $\rho^m \leqq b = 1+\sup_\Omega \rho_0$. Thus we obtain

$$\left(1-\frac{db}{\mu}\right)\int_\Omega \rho^m\left|\frac{\partial u^m}{\partial t}\right|^2 dx + \mu\frac{d}{dt}\int_\Omega |\nabla u^m|^2\,dx + \frac{d\mu}{2}\int_\Omega |P\Delta u^m|^2\,dx$$
$$\leqq \left(b+\frac{db^2}{\mu}\right)\int_\Omega |(u^m\cdot\nabla)u^m - f|^2\,dx.$$

Choosing $d = \mu/2b$ and using Lemma 15(i) it follows that

$$\frac{1}{2}\int_\Omega \rho^m\left|\frac{\partial u^m}{\partial t}\right|^2 dx + \mu\frac{d}{dt}\int_\Omega |\nabla u^m|^2\,dx + 2\varepsilon\int_\Omega |\Delta u^m|^2\,dx$$
$$\leqq 3b\left(\int_\Omega |(u^m\cdot\nabla)u^m|^2\,dx + \int_\Omega |f|^2\,dx\right). \tag{23}$$

By Lemma 15(ii) and by $\alpha\beta \leqq \lambda\alpha^{4/3} + C_\lambda\beta^4$, there holds

$$\int_\Omega |(u^m\cdot\nabla)u^m|^2\,dx \leqq (|u^m|_{(L^\infty(\Omega))^3})^2\int_\Omega |\nabla u^m|^2\,dx$$
$$\leqq C\left(\int_\Omega |\Delta u^m|^2\,dx\right)^{3/4}\left(\int_\Omega |\nabla u^m|^2\,dx\right)^{5/4} \tag{24}$$
$$\leqq \frac{\varepsilon}{3b}\int_\Omega |\Delta u^m|^2\,dx + \delta\left(\int_\Omega |\nabla u^m|^2\,dx\right)^5.$$

Carrying back this estimate in (23), we obtain

$$\frac{1}{2}\int_\Omega \rho^m \left|\frac{\partial u^m}{\partial t}\right|^2 dx + \mu \frac{d}{dt}\int_\Omega |\nabla u^m|^2\, dx + \varepsilon \int_\Omega |\Delta u^m|^2\, dx$$
$$\leqq 3b\delta\left(\int_\Omega |\nabla u^m|^2\, dx\right)^5 + 3b\int_\Omega |f|^2\, dx. \tag{25}$$

Estimates on u^m and u. Let $u_{0,i}$ be the coordinates of u_0 in the basis (w_i), that is, $u_0 = \Sigma_i u_{0,i} w_i$. Choose

$$u_0^m = \sum_{1 \leqq i \leqq m} u_{0,i} w_i.$$

Since this basis is orthogonal in H, $u_0^m \to u_0$ in H, and then hypothesis (16) is satisfied. Moreover, since $(\nabla w_i, \nabla w_j)_\Omega = 0$ if $i \neq j$, there holds

$$\int_\Omega |\nabla u_0^m|^2\, dx = \sum_{1 \leqq i \leqq m} |u_{0,i}|^2 \int_\Omega |\nabla w_i|^2\, dx \leqq \int_\Omega |\nabla u_0|^2\, dx.$$

Then (25) implies, by Lemma 6, that there exists $T_* > 0$ independent of m such that, for all $t \leqq T_*$,

$$\left(\int_\Omega |\nabla u^m|^2\, dx\right)(t) \leqq \int_\Omega |\nabla u_0|^2\, dx + 1.$$

Thus u^m is bounded in $L^\infty(0, T_*; V)$. Moreover, by integrating (25) on $[0, T_*]$, Δu^m is bounded in $L^2(0, T_*; (L^2(\Omega))^3)$. Thus

$$u^m \text{ is bounded in } L^2(0, T_*; (H^2(\Omega))^3) \cap L^\infty(0, T_*; V)$$

and its limit satisfies

$$u \in L^\infty(0, T_*; V) \cap L^2(0, T_*; (H^2(\Omega))^3).$$

Supplementary properties on u and ρ. (a) By Sobolev's and Riesz's theorems u^m is bounded in $L^\infty(0, T_*; (L^6(\Omega))^3) \cap L^{4-\varepsilon}(0, T_*; (L^\infty(\Omega))^3)$ for all $\varepsilon > 0$. Thus $\partial \rho^m / \partial t = -\nabla \cdot (\rho^m u^m)$ satisfies

$$\frac{\partial \rho^m}{\partial t} \text{ is bounded in } L^\infty(0, T_*; W^{-1,6}(\Omega)) \cap L^{4-\varepsilon}(0, T_*; W^{-1,\infty}(\Omega))$$

and ρ, which is the limit of ρ^m, satisfies

$$\frac{\partial \rho}{\partial t} \in L^\infty(0, T_*; W^{-1,6}(\Omega)) \cap L^{4-\varepsilon}(0, T_*; W^{-1,\infty}(\Omega)).$$

(b) By Lemma 3(iii) the product is continuous from $W^{1,6} \times W^{-1,6}$ into $W^{-1,6}$. Thus, since u^m is bounded in $L^2(0, T_*; (W^{1,6}(\Omega))^3)$,

$$u^m \frac{\partial \rho^m}{\partial t} \text{ is bounded in } L^2(0, T_*; (W^{-1,6}(\Omega))^3).$$

On the other hand, the integration of (25) on $[0, T_*]$ gives

$$(\rho^m)^{1/2} \frac{\partial u^m}{\partial t} \text{ is bounded in } L^2(0, T_*; (L^2(\Omega))^3). \tag{26}$$

Thus $\rho^m \partial u^m/\partial t$ is bounded in the same space, which is included in $L^2(0, T_*; (W^{-1,6}(\Omega))^3)$. By addition $\partial \rho^m u^m/\partial t$ is bounded in this space, thus

$$\frac{\partial \rho u}{\partial t} \in L^2(0, T_*; (W^{-1,6}(\Omega))^3).$$

(c) It follows that ρu is continuous on $[0, T_*]$ into $(W^{-1,6}(\Omega))^3$. Since ρu is bounded on $[0, T_*]$ into $(L^6(\Omega))^3$, it is continuous into $(L^6(\Omega))^3$ weak and therefore into $(W^{-1,\infty}(\Omega))^3$.

(d) The distributions $\nabla \cdot (\rho u u)$, Δu, f, and $\partial \rho u/\partial t$ lie in $L^2(0, T_*; (W^{-1,6}(\Omega))^3)$. Then by (1), ∇p lies in the same space, and this implies that $p \in L^2(0, T_*; L^6(\Omega))$.

(e) By Proposition 13(i) the intermediate initial condition is satisfied. This ends the proof of part (i) of Theorem 14.

Proof of part (ii) *of Theorem* 14. (a) By (i), $\partial \rho/\partial t \in L^\infty(0, T_*; W^{-1,6}(\Omega))$, and by Theorem 9, $\rho(0) = \rho_0$. Then, for $0 \leqq t \leqq T_*$,

$$|\rho_0 - \rho(t)|_{W^{-1,6}(\Omega)} \leqq ct.$$

Since $u(t)$ is bounded in $(H^1(\Omega))^3$ for $0 \leqq t \leqq T_*$ it follows, by Lemma 3(iii), that

$$|(\rho_0 - \rho(t))u(t)|_{(W^{-1,3}(\Omega))^3} \leqq ct.$$

Thus $t \to (\rho_0 u(t) - \rho(t)u(t))$ is continuous at zero into $(W^{-1,3}(\Omega))^3$. Moreover, ρu is continuous into this space by (i) of Theorem 14. Therefore $t \to \rho_0 u(t)$ is continuous at zero into $(W^{-1,3}(\Omega))^3$.

Moreover, $\rho_0 u(t)$ is bounded on $[0, T_*]$ into $(L^6(\Omega))^3$. Thus $t \to \rho_0 u(t)$ is continuous at zero into $(L^6(\Omega))^3$ weak. By assumption $1/\rho_0$ is in $L^{6/5}(\Omega)$, thus $t \to u(t)$ is continuous at zero into $(L^1(\Omega))^3$ weak.

Moreover, $u(t)$ is bounded on $[0, T_*]$ into V. Thus $t \to u(t)$ is continuous at zero into $(H_0^1(\Omega))^3$ weak and into $(L^2(\Omega))^3$ strong. Thus $t \to u(t)$ is continuous at zero into H.

(b) In (ii) of Proposition 13 we have proved that

$$(\rho u)(0) = \rho(0)u(0) = \rho_0 u_0$$

when $t \to u(t)$ is continuous on $[0, T]$ into H. But the proof used solely the continuity at $t = 0$. Thus this property is still satisfied here. Multiplying by $1/\rho_0$, it gives $u(0) = u_0$.

Proof of part (iii) *of Theorem* 14. (a) By (26)

$$(\rho^m)^{1/2} \frac{\partial u^m}{\partial t} \text{ is bounded in } L^2(0, T_*; (L^2(\Omega))^3).$$

Moreover, u^m, ρ^m satisfy all the estimates obtained in the proof of Theorem 9. In particular, (17) is satisfied, and thus $\rho^m \geqq \inf \rho_0$. Therefore, when $\inf \rho_0 > 0$,

$$\frac{\partial u^m}{\partial t} \text{ is bounded in } L^2(0, T_*; (L^2(\Omega))^3).$$

Moreover, $\partial u^m/\partial t \in C([0, T]; V)$. Thus it is bounded in $L^2(0, T_*; H)$ and its limit satisfies

$$\frac{\partial u}{\partial t} \in L^2(0, T_*; H).$$

(b) This property and $u \in L^2(0, T_*; (H^2(\Omega))^3)$ imply, by Theorem 3.1 of [12, p. 68], that

$$u \in C([0, T_*]; (H^1(\Omega))^3).$$

Since $u \in L^\infty(0, T_*; V)$ there holds $u \in C([0, T_*], V)$. □

Remark. The estimate (25) is due to Kim [7]. His solution is only local since he does not have the estimates of Theorem 9. He obtains neither the pressure nor the initial condition on ρu.

He supposes that $f \in L^\infty(0, T; (L^2(\Omega))^3)$ to conclude, from (25), that $\int_\Omega |\nabla u^m|^2 \, dx$ is locally bounded. We obtain the same result for $f \in L^2(0, T; (L^2(\Omega))^3)$ by using Lemma 6.

Remark. The local existence of a more regular solution is proved by Ladyzhenskaya and Solonnikov [18]. Assuming

$$u_0 \in (W^{2-2/q,q}(\Omega))^3 \cap V, \quad \rho_0 \in C^1(\bar{\Omega}), \quad \inf_\Omega \rho_0 > 0, \quad f \in (L^q(Q))^3$$

with $q > 3$, they proved the existence of $T_* > 0$ such that for $T = T_*$ there exists a solution u, ρ, p of the exact equations such that

$$u \in L^q(0, T; (W^{2,q}(\Omega))^3 \cap V), \quad \frac{\partial u}{\partial t} \in (L^q(Q))^3, \quad p \in W^{1,q}(\Omega), \quad \rho \in C^1(\bar{Q}).$$

In addition they prove the uniqueness in this class and the global existence, that is, the existence for any given T, for small enough data.

9. Space dimension 2. Now let Ω be a subset of $\mathbb{R}^2$ instead of $\mathbb{R}^3$; we still suppose Ω open, bounded with a Lipschitz boundary. All the previous results remain true, and some may be improved.

At first for the solution with a weak initial condition given in § 5, some of the estimates are satisfied with better coefficients.

PROPOSITION 16. *For $\Omega \subset \mathbb{R}^2$ the solution given in Theorem 9 satisfies, for all $q < \infty$, and for all $\varepsilon > 0$,*

$$\rho u \in L^2(0, T; (L^q(\Omega))^3) \cap N^{1/4,2}(0, T; (W^{-1,2-\varepsilon}(\Omega))^3),$$

$$\frac{\partial \rho}{\partial t} \in L^2(0, T; W^{-1,q}(\Omega)),$$

$$\left| \frac{\partial}{\partial t} \int_\Omega \rho u \cdot v \, dx \right| \leqq g |v|_V \quad \forall v \in V \quad \text{where } g \in L^{2-\varepsilon}(0, T).$$

Proof. The proof is the same as that for Propositions 7 and 8, with the imbeddings $H^1(\Omega) \subset L^q(\Omega)$, for all $q < \infty$ and $H^1(\Omega) \times H^{-1}(\Omega) \subset W^{-1,2-\varepsilon}(\Omega)$ for all $\varepsilon > 0$.

The solution with a strong initial condition is obtained with a weaker assumption on ρ_0 than in § 8. Moreover if ρ_0 is bounded from below, the solution is globally regular, that is, regular on all of $[0, T]$. More precisely, we have Theorem 17.

THEOREM 17. (i) *For $\Omega \subset \mathbb{R}^2$ the solution given in Theorem 14 satisfies, for all $q < \infty$,*

$$p \in L^2(0, T_*; (L^q(\Omega))^2),$$

$$\rho u \in L^\infty(0, T_*; (L^q(\Omega))^2), \qquad \frac{\partial \rho u}{\partial t} \in L^2(0, T_*; (W^{-1,q}(\Omega))^2),$$

$$\frac{\partial \rho}{\partial t} \in L^\infty(0, T_*; W^{-1,q}(\Omega)) \cap L^q(0, T_*; W^{-1,\infty}(\Omega)).$$

(ii) *In* (ii) *of Theorem* 14, *the assumption* $1/\rho_0 \in L^{6/5}(\Omega)$ *may be replaced by*

$$1/\rho_0 \in L^r(\Omega) \quad \text{where } r > 1.$$

(iii) *If in addition* $\inf \rho_0 > 0$, *then all the results of the present theorem and of Theorem* 17 *are satisfied with* $T_* = T$.

Remark. The results of part (iii) are due to Antonzev and Kajikov [2].

Remark. The existence and uniqueness have been proved by Kajikov [6] in a class of smooth solutions $u \in C^{2+\beta,1+\beta/2}(\bar{\Omega} \times [0, T])$, $\rho \in C^1(\bar{\Omega} \times [0, T])$.

Proof of Theorem 17. The proof of (i) is the same as for Theorem 14, by now using $H^1(\Omega) \subset L^q(\Omega)$, for all $q < \infty$.

(ii) Now $\rho_0 u(t)$ is bounded on $[0, T_*]$ into $(L^q(\Omega))^2$, for all $q < \infty$. Thus the proof of Theorem 14(ii) gives us that $t \to \rho_0 u(t)$ is continuous at zero into $(L^q(\Omega))^2$ weak. Thus, choosing $1/q + 1/r = 1$, the assumption is enough to conclude that $t \to u(t)$ is continuous at zero into $(L^1(\Omega))^2$ weak.

(iii) For $v \in H^2(\Omega)$ there holds

$$|v|_{L^4(\Omega)} \leqq c|v|^{1/2}_{L^2(\Omega)}|v|^{1/2}_{H^1(\Omega)},$$

$$|\nabla v|_{(L^4(\Omega))^2} \leqq c|v|^{1/2}_{H^1(\Omega)}|v|^{1/2}_{H^2(\Omega)} \leqq c|v|^{1/4}_{L^2(\Omega)}|v|^{3/4}_{H^2(\Omega)}.$$

Thus in the proof of Theorem 14, estimate (24) may be replaced by

$$\int_\Omega |(u^m \cdot \nabla)u^m|^2 \, dx \leqq \left(\int_\Omega |u^m|^4 \, dx\right)^{1/2} \left(\int_\Omega |\nabla u^m|^4 \, dx\right)^{1/2}$$

$$\leqq c|u^m|_{(L^2(\Omega))^2}|u^m|^2_{(H^1(\Omega))^2}|u^m|_{(H^2(\Omega))^2}$$

$$\leqq \varepsilon/3b \int_\Omega |\Delta u^m|^2 \, dx + \delta\left(\int_\Omega |\nabla u^m|^2 \, dx\right)^2.$$

Now (25) yields

$$\frac{1}{2}\int_\Omega \rho^m \left|\frac{\partial u^m}{\partial t}\right|^2 dx + \mu \frac{d}{dt}\int_\Omega |\nabla u^m|^2 \, dx + \varepsilon \int_\Omega |\Delta u^m|^2 \, dx$$

$$\leqq 3b\delta k^m \int_\Omega |\nabla u^m|^2 \, dx + 3b \int_\Omega |f|^2 \, dx$$

where $k^m = \int_\Omega |\nabla u^m|^2 \, dx$ is bounded in $L^1(0, T)$.

Thus u^m are bounded in $L^\infty(0, T; (H^1(\Omega))^2) \cap L^2(0, T; (H^2(\Omega))^2)$ by Gronwall's lemma. From this we can choose $T_* = T$ in the third step of the proof of Theorem 14(i). □

Acknowledgments. The author is indebted to J. C. Saut, who encouraged him to work on these questions, and to the referee for many improvements, in particular, in the case of space dimension 2.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] S. A. ANTONZEV AND A. V. KAJIKOV, *Mathematical study of flows of nonhomogeneous fluids*, Lectures at the University of Novosibirsk, Novosibirsk, U.S.S.R., 1973. (In Russian.)

[3] L. CATTABRIGA, *Su un problema al contorno relativo al sistema di equazioni di Stokes*, Rend. Sem. Mat. Padova, 31 (1961), pp. 308–340.

[4] V. GIRAULT AND P. A. RAVIART, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, New York, 1986.
[5] A. V. KAJIKOV, *Resolution of boundary value problems for non homogeneous viscous fluids*, Dokl. Akad. Nauk., 216 (1974), pp. 1008-1010.
[6] ———, *Mathematical study of flows of non-homogeneous fluids*, In Seminar on Numerical Methods in Mechanics (Sem. N. N. Yanenko, Novosibirsk), Part II, 1975, pp. 65-76.
[7] U. J. KIM, *Weak solutions of an initial boundary value problem for an incompressible viscous fluid with nonnegative density*, SIAM J. Math. Anal., 18 (1987), pp. 89-96.
[8] O. LADYZHENSKAYA, *The Mathematical Theory of Viscous Incompressible Fluids*, Gordon and Breach, New York, 1969.
[9] J. L. LIONS, *On some questions in boundary value problems of mathematical physics*, Federal University, Rio de Janeiro, Brazil, 1977.
[10] ———, *On some problems connected with Navier-Stokes equations in nonlinear evolution equations*, M. C. Crandall, ed., Academic Press, New York 1978.
[11] ———, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.
[12] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Vol. 1, Dunod, Paris, 1968.
[13] J. SIMON, *Ecoulement d'un fluide non homogène avec une densité initiale s'annulant*, C. R. Acad. Sci. Paris Ser. A, 15 (1978), pp. 1009-1012.
[14] ———, *Compact sets in the space $L^p(0, T; B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65-96.
[15] ———, *Sobolev, Besov and Nikolskii fractional spaces: imbeddings and comparisons for vector valued spaces on an interval*, Ann. Mat. Pura Appl., to appear.
[16] ———, *Equations fortes vérifiées par les solutions faibles des équations de Navier-Stokes*, Mémoire d'habilitation à diriger des recherches, Université Paris 6, Paris, France, 1988.
[17] L. TARTAR, *Topics in nonlinear analysis*, Publ. Math. Orsay, Orsay, France, 1978.
[18] O. LADYZHENSKAYA AND V. A. SOLONNIKOV, *Unique solvability of an initial and boundary value problem for viscous incompressible non-homogeneous fluids*, J. Soviet. Math., 9 (1978), pp. 697-749.
[19] M. PADULA, *An existence theorem for non-homogeneous incompressible fluids*, Rend. Circ. Mat. Palermo (4), 31 (1982), pp. 119-124.
[20] H. OKAMOTO, *On the equation of nonstationary stratified fluid motion: uniqueness and existence of solutions*, J. Fac. Sci. Univ. Tokyo, Sect. 1A Math., 30 (1984), pp. 615-643.

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1118-1138, September 1990
© 1990 Society for Industrial and Applied Mathematics

A DYNAMIC FREE-BOUNDARY PROBLEM IN PLASMA PHYSICS*

YOSHIKAZU GIGA† AND ZENSHO YOSHIDA‡

Abstract. A viscous incompressible inhomogeneous plasma in a bounded domain in $\mathbf{R}^N$ ($N=2,3$) is considered. The resistivity is discontinuous and the set of discontinuities corresponds to the interface (free boundary) of two plasmas having different temperatures. The dynamics of the free boundary is studied. A reasonable model system is a parabolic system with discontinuous coefficient coupled with a transport equation for the resistivity. No assumptions are imposed on the regularity of the initial interface. A priori estimates are derived for the resistivity in the class of functions of bounded variations, and global-in-time weak solutions are constructed for the model equations such that the area of the free boundary does not blow up in finite time provided that either $N=2$, or the difference of resistivity is small (for $N=3$). The fixed boundary $\partial\Omega$ surrounding the total plasma is assumed to be perfectly conductive and adherent.

Key words. plasma, free boundary, parabolic system, discontinuous coefficients, transport equations

AMS(MOS) subject classifications. 35F10, 35K20, 35R05, 35R35, 76W05

1. Introduction. This paper studies the dynamics of the interface (free boundary) of two incompressible viscous plasmas—high and low temperature plasmas in a smoothly bounded domain Ω in $\mathbf{R}^N$ ($N=2,3$). Each plasma has different electric resistivity $\eta_L > \eta_H > 0$ which are assumed to be constants for simplicity. A large η_L (small η_H) corresponds to low- (high-) temperature plasma. We consider a step function η with two values η_L and η_H. The region occupied with low (high) plasma corresponds to the place where η takes the value η_L (η_H). The interface corresponds to a jump discontinuity of η. We study the dynamics of η instead of the interface itself. The kinematic viscosity is assumed to be constant on whole Ω.

Plasma confinement experiments, for example, in controlled fusion researches, produce high-temperature plasma that are necessarily detached from the boundary wall materials. A high-temperature plasma is generally surrounded by a low temperature plasma.

The evolution of η is described by a convection transport equation:

$$\partial_t \eta + (v \cdot \nabla)\eta = 0 \quad \text{in } \Omega,$$

where v is the fluid velocity. This v is governed by momentum equations of the incompressible Navier–Stokes system. The momentum equations have the magnetic force term, which couples the equations with the magnetic induction-diffusion equations which are essentially Ohm's law. The discontinuous resistivity function η appears in the magnetic diffusion term. The boundary $\partial\Omega$ is assumed to be perfectly conductive and adherent.

This self-consistent model describes many interesting phenomena of plasmas. Some variety of free-boundary instabilities have been studied by linear stability analysis, and nonlinear numerical simulations also have been developed (for example, see [1]). For constant η a global (in time) regular solution is known to exist for arbitrary

* Received by the editors November 7, 1988; accepted for publication (in revised form) October 20, 1989.

† Department of Mathematics, Hokkaido University, Sapporo 060, Japan. The work of this author was partially supported by Grants-in-Aid for Scientific Research 62460001, 62740071, 63740064, The Japan Ministry of Education, Science, and Culture.

‡ Department of Nuclear Engineering, University of Tokyo, Tokyo 113, Japan, and Institute for Fusion Theory, Hiroshima University, Hiroshima 730, Japan. The work of this author was partially supported by Grant-in-Aid for Scientific Research 63050040, The Japan Ministry of Education, Science, and Culture.

(respectively, small) initial data when $N=2$ ($N=3$) (cf. [20]). In the present work we consider a perturbation around an equilibrium and drop quadratic terms of perturbations of the magnetic field and the fluid velocity, however, keeping $(v\cdot\nabla)\eta$ in the transport equation. We also assume, just for simplicity, that the equilibrium is current-free. For this model system, we construct a weak global solution such that the area of the free boundary does not blow up in a finite time provided that either the space dimension $N=2$ or $\eta_L/\eta_H \sim 1$ for $N=3$. We do not even need to assume that η is a locally constant function to construct a weak global solution. We consider the transport equation in BV, the space of functions of bounded variations. If η is in BV with two values η_L and η_H, the (essential) total variation in space variables $\|\nabla\eta\|_1(t)$ equals $\eta_L-\eta_H$ times the area of the interface at the time t. Here the interface should be defined as the (approximate) jump discontinuity of η [5], [10], [19]. In this paper we do not assume that initial interface is smooth, but only assume initial $\|\nabla\eta\|_1(0)$ is finite.

To construct the solution of our system, we appeal to the Schauder fixed-point theorem. The crucial step is to obtain various a priori estimates. We need a regularity of v to get a priori estimates for η in BV space. The regularity of v follows from the regularity of the magnetic field b. However, since the magnetic diffusion coefficient η is discontinuous, we do not expect much regularity for b. Although it is easy to show that ∇b is a space-time L^2 function, by an energy equality, this does not give enough regularity for v. We will prove ∇b is a space-time L^r function for some $r>N$ by a perturbation argument. Here we need to assume that $N=2$, or $\eta_L/\eta_H\sim 1$ for $N=3$. Such a perturbation argument is found in Campanato [4] for Dirichlet problems. Since our boundary condition for b is Neumann type (see § 2)

$$n\times \operatorname{rot} b=0, \qquad n\cdot b=0 \quad \text{on } \partial\Omega,$$

his theory cannot be applied directly to our problem. We are forced to use an approximation argument, since the trace $n\times\operatorname{rot} b$ is not well defined for b, when ∇b is only a (space) L^2 function.

This paper is organized as follows. In § 2, we formulate our problem including a two-dimensional version. In § 3, we state our main results, and sketch our scheme to construct solutions. In § 4, we solve the transport equation in BV space assuming that v is a given function. Section 5 is devoted to deriving a priori L^r estimates for parabolic systems with discontinuous coefficients when the boundary condition is a Neumann type. In § 6, we construct a weak solution by applying the Schauder fixed-point theorem.

2. Formulation of the problem. We consider an incompressible plasma (a current-carrying fluid) with inhomogeneous resistivity. The plasma is contained in a smoothly bounded domain Ω in $\mathbf{R}^3$. The boundary $\partial\Omega$ is assumed to be perfectly conductive and adherent. We begin by formulating a model equation when the space dimension N equals 3. The dynamic state of the system is characterized by the following functions:

$b(x,t)$: magnetic flux density (three vector),
$v(x,t)$: fluid velocity (three vector),
$p(x,t)$: pressure (scalar),
$\eta(x,t)$: resistivity (scalar).

We here consider an incompressible viscous plasma whose kinematic viscosity ν and density ρ are positive constants. The dynamics of the plasma is described by the following system of partial differential equations:

$$\partial_t b=-\operatorname{rot}\eta\operatorname{rot} b+\operatorname{rot}(v\times b), \qquad \operatorname{div} b=0, \tag{2.1}$$

$$\partial_t v = \nu \Delta v - (v \cdot \nabla) v - b \times \operatorname{rot} b/(\mu_0 \rho) - \operatorname{grad} p/\rho, \qquad \operatorname{div} v = 0, \tag{2.2}$$

$$\partial_t \eta + (v \cdot \nabla)\eta = 0, \tag{2.3}$$

where $\partial_t = \partial/\partial t$, rot η rot $b = \operatorname{rot}(\eta \operatorname{rot} b)$, and μ_0 represents the vacuum permeability. Equation (2.1) is essentially Ohm's law. Equation (2.2) is the Navier–Stokes system with the magnetic-force term $-b \times \operatorname{rot} b$; here grad p is the gradient field of p and

$$(v \cdot \nabla) = \sum_{i=1}^{N} v_i \frac{\partial}{\partial x_i},$$

where $v = (v_1, \cdots, v_N)$. Equation (2.3) is the transport equation for the resistivity η. Since the boundary is perfectly conductive and adherent, the boundary conditions we impose are

$$n \cdot b = g(x), \quad n \times \operatorname{rot} b = 0, \quad v = 0 \quad \text{on } \partial\Omega,$$

where $g(x)$ is a given time-independent function with vanishing mean value and n is the exterior unit normal vector to $\partial\Omega$.

To avoid other difficulties, we consider small perturbations b, v, and p around a current-free equilibrium B, V, P such that

$$\begin{aligned} &\operatorname{rot} B = 0, \quad \operatorname{div} B = 0, \quad V = 0, \quad P = 0 \quad \text{in } \Omega, \\ &n \cdot B = g(x) \quad \text{on } \partial\Omega. \end{aligned} \tag{2.4}$$

We plug $b + B$, $v + V$ and $p + P$ in (2.1)–(2.3), and neglect quadratic terms of perturbations b, v, p. System (2.1)–(2.3) now yields

$$\partial_t b = -\operatorname{rot} \eta \operatorname{rot} b + \operatorname{rot}(v \times B), \qquad \operatorname{div} b = 0, \tag{2.5}$$

$$\partial_t v = \Delta v - B \times \operatorname{rot} b - \operatorname{grad} p, \qquad \operatorname{div} v = 0, \tag{2.6}$$

$$\partial_t \eta + (v \cdot \nabla)\eta = 0, \tag{2.7}$$

and on $\partial\Omega$

$$n \cdot b = 0, \quad n \times \operatorname{rot} b = 0, \quad v = 0, \tag{2.8}$$

where we set $\nu = \mu_0 = \rho = 1$ for simplicity of notation; our existence result in § 3 is still valid without assuming $\nu = \mu_0 = \rho = 1$. The boundary condition for b at the interface (jump discontinuity of η) is implicit in (2.5):

$$[\eta n \times \operatorname{rot} b] = 0,$$

where $[f]$ is the jump of a function f at the free boundary.

In this paper, we also consider a two-dimensional version of the problem. This is obtained by considering an ignorable coordinate, say x_3 of $x = (x_1, x_2, x_3)$. However, it is necessary to be careful in definitions of the rot operators and the vector products. For a two-vector function $w = (w_1(x_1, x_2), w_2(x_1, x_2))$, we define

$$\operatorname{rot} w = \frac{\partial w_2}{\partial x_1} - \frac{\partial w_1}{\partial x_2}$$

which gives a scalar function. We also define, for a scalar function ψ,

$$\operatorname{rot}^* \psi = \left(\frac{\partial \psi}{\partial x_2}, -\frac{\partial \psi}{\partial x_1} \right)$$

which gives a two-vector function. For a three-vector function u, we define

$$\operatorname{rot}^* u = \operatorname{rot} u,$$

where rot is the usual rotational operator. We use the notation of the exterior product instead of the vector product. For two-vectors $u=(u_1, u_2)$ and $v=(v_1, v_2)$ we write

$$u\Lambda v = u_1 v_2 - u_2 v_1.$$

For scalar ψ, we define a two-vector by

$$u\Lambda^*\psi = (u_2\psi, -u_1\psi).$$

For three-vectors u and v, we define

$$u\Lambda v = u\Lambda^* v = u \times v.$$

Using these conventions of notation, we now write the model equations (2.5)-(2.8) in a general form which is valid for both two and three dimensions:

$$\partial_t b = -\mathrm{rot}^*\eta \ \mathrm{rot}\, b + \mathrm{rot}^*(v\Lambda B), \qquad \mathrm{div}\, b = 0 \quad \text{in } \Omega, \tag{2.9}$$

$$\partial_t v = \Delta v - B\Lambda^* \ \mathrm{rot}\, b - \mathrm{grad}\, p, \qquad \mathrm{div}\, v = 0 \quad \text{in } \Omega, \tag{2.10}$$

$$\partial_t \eta + (v\cdot\nabla)\eta = 0 \quad \text{in } \Omega, \tag{2.11}$$

$$n\cdot b = 0, \quad n\Lambda^* \ \mathrm{rot}\, b = 0, \quad v = 0 \quad \text{on } \partial\Omega, \tag{2.12}$$

where Ω is in $\mathbf{R}^N$ and $N=2$ or 3.

3. Main results and outline of the proof. Our goal is to prove the existence of global-in-time solution of the initial value problem for (2.9)-(2.12) even when initial resistivity η_0 is discontinuous. We first list our assumptions on initial data. We assume that initial resistivity $\eta_0 = \eta_o(x)$ satisfies

$$\eta_0 \in \mathrm{BV}(\Omega) \quad \text{and} \quad 0 < \alpha \leqq \eta_0 \leqq \beta \quad \text{in } \Omega, \tag{3.1}$$

where α and β are constants. Here $\mathrm{BV}(\Omega)$ is the set of functions of bounded variations (see [10], [19]), i.e.,

$$\mathrm{BV}(\Omega) = \{\eta;\ \eta \in L^1(\Omega), \nabla\eta \in M(\Omega)\},$$

where $M(\Omega)$ is the set of finite Radon measures on Ω and ∇ is a (distributional) gradient. Here and hereafter we do not distinguish between spaces of vector-valued and scalar-valued functions. For the initial velocity v_0, we assume

$$\begin{aligned} &v_0, \nabla^2 v_0 \in L^q(\Omega), \qquad \mathrm{div}\, v_0 = 0 \quad \text{in } \Omega, \\ &v_0 = 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{3.2}$$

with $q>1$, where $L^q(\Omega)$ is a usual Lebesgue space and ∇^2 represents any spatial (distributional) derivative of the second order. The initial magnetic field b_0 satisfies

$$\begin{aligned} &b_0, \mathrm{rot}\, b_0, \quad \nabla\, \mathrm{rot}\, b_0 \in L^q(\Omega), \quad \mathrm{div}\, b_0 = 0 \quad \text{in } \Omega, \\ &n\cdot b_0 = 0, \quad n\Lambda^* \ \mathrm{rot}\, b_0 = 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{3.3}$$

The trace $n\cdot b_0$ on $\partial\Omega$ is well defined because $\mathrm{div}\, b_0 = 0$ and $b_0 \in L^q(\Omega)$ (see [7]). Since our $\eta(x, t)$ may have jump discontinuity, we should interpret (2.9)-(2.12) in some weak sense.

If v is smooth, for η we expect, from (2.11) and (3.1), that

$$\begin{aligned} &\eta(x, t) \in \mathrm{BV}(Q_T) \cap C([0, T); L^1(\Omega)), \\ &\nabla\eta \in L^\infty(0, T; M(\Omega)), \\ &\partial_t \eta \in L^1(0, T; M(\Omega)), \end{aligned} \tag{3.4}$$

where $Q_T = \Omega \times [0, T)$; $f \in L^r(0, T; X)$ means that the mapping from t to $f(\cdot, t)$ is L^r with value in X, and $f \in C([0, T); X)$ means that the mapping from t to $f(\cdot, t)$ is continuous with value in X, where X is a Banach space. Equation (2.11) can be understood as an equality of measures on Q_T provided that v is smooth. The initial condition $\eta|_{t=0} = \eta_0$ is well defined since $\eta \in C([0, T); L^1(\Omega))$.

If b and B are not singular, it is natural to expect that our v satisfies

$$v(x, t) \in W_r^{2,1}(Q_T), \tag{3.5}$$

viz., $v, \partial_t v, \nabla^2 v \in L^r(Q_T)$ for some r, since (2.10) is parabolic. The traces of v on $t = 0$ and on $\partial\Omega$ are well defined for such functions. Both sides of (2.10) can be understood as an equality of L^r functions.

Since η is discontinuous, for b we only expect

$$b(x, t) \in C([0, T); L^2_\sigma(\Omega)) \cap L^2(0, T; W), \quad W = H^1(\Omega) \cap L^2_\sigma(\Omega), \tag{3.6}$$

where $H^1(\Omega)$ is the Sobolev space of order 1, and

$$L^2_\sigma(\Omega) = \{u \in L^2(\Omega); \operatorname{div} u = 0 \text{ in } \Omega,\ n \cdot u = 0 \text{ on } \partial\Omega\}. \tag{3.7}$$

We need a weak formulation of (2.9) so that we can interpret boundary condition $n\Lambda^* \operatorname{rot} b = 0$. We define a bilinear form

$$a(t, \psi, \phi) = \int_\Omega [\eta(x, t) \operatorname{rot} \psi \cdot \operatorname{rot} \phi]\, dx. \tag{3.8}$$

A weak formulation of (2.9) with $n\Lambda^* \operatorname{rot} b = 0$ and $b|_{t=0} = b_0$ is, for any $T > 0$,

$$\begin{aligned} &\int_0^T \int_\Omega [b \cdot \partial_t \phi]\, dx\, dt + \int_\Omega [b_0 \phi(x, 0)]\, dx \\ &\quad = -\int_0^T \left\{ a(t, b, \phi) + \int_\Omega [\operatorname{rot}^*(v \Lambda B) \cdot \phi]\, dx \right\} dt \end{aligned} \tag{3.9}$$

holding for every $\phi(x, t) \in C_0^1(\bar{\Omega} \times [0, T))$, i.e., ϕ is C^1 with a compact support in $\bar{\Omega} \times [0, T)$. Since the condition $\operatorname{div} b = 0$ and $n \cdot b = 0$ is included in (3.6), the weak formulation (3.9) is formally equivalent to (2.9), (2.12) with $b|_{t=0} = b_0$. We thus see that system (2.9)-(2.12) with initial conditions

$$b|_{t=0} = b_0, \quad v|_{t=0} = v_0, \quad \eta|_{t=0} = \eta_0$$

is formally equivalent to (3.9), (2.10), (2.11) with

$$v = 0 \quad \text{on } \partial\Omega, \tag{3.10}$$

$$v|_{t=0} = v_0 \quad \text{in } \Omega, \tag{3.11}$$

$$\eta|_{t=0} = \eta_0 \quad \text{in } \Omega, \tag{3.12}$$

under the condition (3.6) on b.

MAIN EXISTENCE THEOREM. *Assume that B is smooth on $\bar{\Omega}$ and satisfies* (2.4). *For positive numbers α and β in* (3.1), *we suppose that β/α is close to* 1 *when the space dimension $N = 3$, and that α and β are arbitrary when $N = 2$. Suppose that initial data (b_0, v_0, η_0) satisfies* (3.1)-(3.3) *with $q > N$. Then for every $T > 0$ there is a function $(b(x, t), v(x, t), \eta(x, t))$ that solves* (3.9), (2.10), (2.11) *with* (3.10)-(3.12) *for some*

$p(x, t)$, and satisfies (3.4)–(3.6) *for some $r > N$ and $\nabla b \in L^r(Q_T)$. It also satisfies the energy equality*

$$\int_\Omega b^2(x, T)\,dx + \int_\Omega v^2(x, T)\,dx + 2\iint_{Q_T} \eta(\mathrm{rot}\, b)^2\,dx\,dt + 2\iint_{Q_T} \eta(\nabla v)^2\,dx\,dt = \int_\Omega b_0^2\,dx + \int_\Omega v_0^2\,dx.$$

Remark. Since $\nabla\eta \in L^\infty(0, T; M(\Omega))$ by (3.4), we see the total variation on Ω, $\|\nabla\eta\|_1(t)$ is bounded on $[0, T)$. If initial η_0 is a two-valued function, this $\|\nabla\eta\|_1(t)$ measures the area of free surface of high-temperature plasma. Our main theorem says the area of the free boundary cannot blow up in finite time at least for our solution.

Remark. The condition rot $B = 0$ in (2.4) is assumed only for simplicity. We may drop this assumption in the Main Existence Theorem; the only change appears in the energy equality.

Remark. Our assumptions on B, b_0, v_0 are not at all optimal. We choose this assumption to avoid using functions spaces with fractional derivatives.

We construct a solution by a fixed-point argument. For a function $\eta_1(x, t)$ satisfying $\eta_1(x, 0) = \eta_0(x)$, we solve (3.9), (2.10) with $\eta = \eta_1$ under (3.10), (3.11) to get a solution (b, v) denoted (b_1, v_1). We then solve the transport equations (2.11) and (3.12) with $v = v_1$ and find a solution denoted η_2. This procedure defines a mapping

$$F: \eta_1 \to \eta_2.$$

We will show the existence of a fixed point of the mapping F using the Schauder fixed-point theorem in appropriate function spaces. A fixed point of F is the desired solution of our problem.

4. Transport equations. To study the evolution of resistivity, we consider a transport equation

$$\partial_t\eta + (v\cdot\nabla)\eta = 0, \tag{4.1}$$

even if initial data is merely a function of bounded variation. We will always assume

$$\mathrm{div}\, v = 0 \quad \text{in } \Omega, \qquad n\cdot v = 0 \quad \text{on } \partial\Omega \quad \text{(for a.e. } t\text{)}; \tag{4.2}$$

the space dimension N is arbitrary in this section. Our $v(x, t)$ is not singular; we assume

$$\begin{aligned} &\nabla v(x, t) \text{ is continuous in } x \in \bar{\Omega} \quad \text{for a.e. } t, \\ &\int_0^T [\|v\|_\infty + \|\nabla v\|_\infty]\,dt < \infty, \end{aligned} \tag{4.3}$$

where $\|\cdot\|_\infty$ is the supremum norm on Ω. Our goal in this section is to prove the existence and uniqueness of solutions of (4.1) with (4.2), (4.3) when η is initially a function of a bounded variation, i.e.,

$$\eta|_{t=0} = \eta_0 \in \mathrm{BV}(\Omega). \tag{4.4}$$

We will also show the continuity of η with respect to v. Since

$$(v\cdot\nabla)\eta = \mathrm{div}\,(v\eta)$$

by (4.2), equation (4.1) can be regarded as a special example of scalar conservation laws. Existence and uniqueness of global (in time) entropy solutions of scalar conservation laws is well known under some regularity assumptions on coefficients when Ω is

$\mathbf{R}^N$ by various methods (e.g., [11], [15], [19]). Similar results are known when Ω has a boundary (see [2]). Usual theories need to assume continuity of the second derivatives of v because (4.2) is not assumed. Although it is not difficult to adjust their arguments to our situation, we give here instead a proof independent of their arguments, because our equation is just linear and we do not need the concept of entropy solutions.

By (4.3), the characteristic equation

$$\frac{dx}{dt} = -v(x, t), \qquad x|_{t=s} = y \in \Omega \tag{4.5}$$

has a unique local solution $x(t, s, y)$. Since $n \cdot v = 0$ by (4.2), the characteristic curve stays in Ω. Since Ω is bounded, x can be extended to a global solution. We note that $x(t, s, y)$ is continuous in (t, s, y), and that $x(t, s, y)$ is a C^1 diffeomorphism of Ω for each t, s, since div $v = 0$ by (4.2). Let $y(t, s, x)$ be its inverse, viz.,

$$\frac{dy}{ds} = v(y, s), \qquad y|_{s=t} = x \in \Omega.$$

We have

$$\left|\frac{\partial y_i}{\partial x_j}\right|(t, s, x) \leqq A \exp\left[\int_s^t a\|\nabla v\|_\infty \, d\tau\right], \qquad 1 \leqq i, \quad j \leqq N, \tag{4.6}$$

with positive constants a and A depending only on N; this follows from

$$\det\left(\frac{\partial y_i}{\partial x_j}\right) = 1, \tag{4.7}$$

since div $v = 0$. We begin with existence of solutions of (4.1)–(4.4) when initial data is C^1.

LEMMA 4.1. *For $\phi_0 \in C^1(\bar{\Omega})$ and $\psi \in C^1(Q_T)$ with $\nabla\psi \in L^1(Q_T)$, we set*

$$\phi(x, t) = \int_0^t \psi(y(t, s, x), s)\, ds + \phi_0(y(t, 0, x)).$$

Then,

$$\nabla\phi \in L^\infty(0, T; L^1(\Omega)), \qquad \partial_t\phi \in L^1(Q_T),$$

and

$$\phi \in C(\bar{Q}_T).$$

Moreover, $\phi(x, t)$ solves

$$\partial_t\phi + (v \cdot \nabla)\phi = \psi \quad \text{in } L^1(Q_T),$$
$$\phi|_{t=0} = \phi_0.$$

Proof. Since $y(t, s, x)$ is continuous in x, t, s, the function ϕ is in $C(\bar{Q}_T)$. We take the (distributional) gradient of ϕ and see

$$\frac{\partial\phi}{\partial x_k} = \sum_{j=1}^N \left[\int_0^t \frac{\partial\psi}{\partial y_j}\frac{\partial y_j}{\partial x_k}\, ds + \frac{\partial\phi_0}{\partial y_j}\frac{\partial y_j}{\partial x_k}\right].$$

Since

$$\int_\Omega |\nabla_y\psi(y, s)|\, dx = \int_\Omega |\nabla_y\psi(y, s)| \cdot \left|\det\left(\frac{\partial x_i}{\partial y_j}\right)\right| dy,$$

we have

$$\|\nabla\phi\|_1(t) \leqq J \cdot K \cdot \left[\int_0^t \|\nabla\psi\|_1\, ds + \|\nabla\phi_0\|_1\right],$$

where $J = \sup_{x\in\Omega}\{\det(\partial x_i/\partial y_j)\}$, K is the supremum of $|\partial y^j/\partial x_i|$ for $t \geqq s \geqq 0$, $x \in \Omega$, $1 \leqq i, j, \leqq N$, and $\|f\|_1 = \int_\Omega |f|\, dx$. By (4.7) we have $J = 1$. We thus conclude from (4.6) that

$$\|\nabla\phi\|_1(t) \leqq A\left[\int_0^t \|\nabla\psi\|_1\, ds + \|\nabla\phi_0\|_1\right] \cdot \exp\int_0^t a\|\nabla v\|_\infty\, ds, \tag{4.8}$$

which yields

$$\nabla\phi \in L^\infty(0, T; L^1(\Omega)). \tag{4.9}$$

A direct calculation shows that

$$\partial_t\phi = -(v\cdot\nabla)\phi + \psi.$$

Since $v \in L^1(0, T; L^\infty(\Omega))$, and $\psi \in L^1(Q_T)$ by boundedness of Q_T, (4.9) shows that $\partial_t\phi$ is in $L^1(Q_T)$. □

We now consider (4.1)–(4.4) when η_0 is merely in $\mathrm{BV}(\Omega)$. The total variation of $\mu \in M(\Omega)$ will be written by $\|\mu\|_1$.

LEMMA 4.2. *Assume that* (4.2) *and* (4.3) *hold. We suppose that* $\eta_0 \in \mathrm{BV}(\Omega) \cap L^\infty(\Omega)$. *Then,*

(i) *The function*

$$\eta(x, t) = \eta_0(y(t, 0, x))$$

is a unique solution of (4.1), (4.4) *on* Q_T *with*

$$\begin{aligned} &\eta \in L^\infty(Q_T), \qquad \partial_t\eta \in L^1(0, T; M(\Omega)), \\ &\nabla\eta \in L^\infty(0, T; M(\Omega)) \quad \text{and} \quad \eta \in C([0, T); L^1(\Omega)). \end{aligned} \tag{4.10}$$

(ii) *Moreover, we have*

$$\|\nabla\eta\|_1(t) \leqq A\|\nabla\eta_0\|_1 \cdot \exp\int_0^T a\|\nabla v\|_\infty\, ds.$$

(iii) *If* $\alpha \leqq \eta_0(x) \leqq \beta$ *for some numbers* α *and* β, *then*

$$\alpha \leqq \eta \leqq \beta \quad \text{on } Q_T.$$

Proof. Since $\nabla v(\cdot, t) \in C(\bar\Omega)$ for a.e. t by (4.3), $(\partial y_i/\partial x_j)$ is continuous in $x \in \bar\Omega$. We thus see

$$\frac{\partial\eta}{\partial x_k} = \sum_{j=1}^N \frac{\partial\eta_0}{\partial y_j}\cdot\frac{\partial y_j}{\partial x_k} \in M(\Omega) \quad \text{for a.e. } t.$$

Similarly to obtain (4.8), (4.6), and (4.7) deduce

$$\|\nabla\eta\|_1(t) \leqq A\|\nabla\eta_0\|_1 \cdot \exp\int_0^T a\|\nabla v\|_\infty\, ds \quad \text{for a.e. } t,$$

which leads to (ii). We thus see $\nabla\eta \in L^\infty(0, T; M(\Omega))$. Taking the (distributional) time derivative, we have

$$\partial_t\eta = -(v\cdot\nabla)\eta.$$

Estimating the right-hand side yields

$$\int_0^T \|\partial_t \eta\|_1 \, ds \leqq \int_0^T \|v\|_\infty \, ds \cdot \sup_{0 \leqq s \leqq T} \|\nabla \eta\|_1(s).$$

We thus observe $\partial_t \eta \in L^1(0, T; M(\Omega))$. The assertion $\eta \in L^\infty(Q_T)$ and (iii) is clear from the definition of η. The property that $\eta \in C([0, T); L^1(\Omega))$ is easy to check.

It remains to prove the uniqueness of the solution of (4.1), (4.4). Suppose that

$$\partial_t \eta + (v \cdot \nabla)\eta = 0, \qquad \eta|_{t=0} = 0,$$

with

$$\partial_t \eta \in L^1(0, T; M(\Omega)), \quad \nabla \eta \in L^\infty(0, T; M(\Omega)), \quad \eta \in L^\infty(Q_T).$$

It is enough to prove $\eta = 0$ on Q_T. We will appeal to a simple duality argument. By Lemma 4.1, for every $\psi \in C^1(\bar{Q}_T)$, there is a function $\phi \in C(\bar{Q}_T)$ that solves

$$\partial_t \phi + (v \cdot \nabla)\phi = \psi \quad \text{in } 0 \leqq t \leqq T, \quad \phi|_{t=T} = 0,$$

with

$$\partial_t \phi \in L^1(Q_T), \qquad \nabla \phi \in L^\infty(0, T; L^1(\Omega)).$$

We multiply ψ by η and integrate over Q_T. By (4.2), integrating by parts yields

$$\begin{aligned}\int_0^T \int_\Omega \psi \eta \, dx \, ds &= \int_0^T \int_\Omega [\partial_t \phi + (v \cdot \nabla)\phi]\eta \, dx \, ds \\ &= -\int_0^T \int_\Omega \phi[\partial_t \eta + (v \cdot \nabla)\eta] \, dx \, ds \\ &\qquad + \int_\Omega \phi(x, T)\eta(x, T) \, dx - \int_\Omega \phi(x, 0)\eta_0(x) \, dx.\end{aligned}$$

This equals zero by definitions of ϕ and η. Since ψ is arbitrary, we conclude that $\eta = 0$. The uniqueness of the solution is now established. □

As an application of Lemma 4.1, we will show the continuity of η with respect to v. Since we have no good uniqueness results as in Lemma 4.2 for

$$\partial_t \eta + (v \cdot \nabla)\eta = \psi$$

when ψ is not in $C^1(Q_T)$, we need to approximate v by smooth functions. In approximating v we should preserve divergence free properties. We prepare a density lemma for our approximations.

LEMMA 4.3. *Assume that* $1 < r < \infty$, *and that* $r > N/2$; *then the space*

$$Q = \{v \in C^\infty(\bar{Q}_T); \operatorname{div} v = 0, \, v|_{\partial\Omega} = 0\}$$

is dense in a Banach space

$$E_r = \left\{ v; |||v|||_r = \int_0^T [\|v\|_\infty + \|\nabla^2 v\|_r](s) \, ds < \infty, \operatorname{div} v = 0, \, v|_{\partial\Omega} = 0 \right\}$$

equipped with norm $|||v|||_r$, *where* $\|f\|_r$ *denotes the* L^r*-norm of* f *in* $L^r(\Omega)$.

Proof. We use a solution of the Stokes equation. As is well known $L^r(\Omega)$ space admits a direct sum decomposition called the Helmholtz decomposition [7]:

$$\begin{aligned}&L^r(\Omega) = L^r_\sigma + G^r, \\ &L^r_\sigma = \{v \in L^r(\Omega); \operatorname{div} v = 0, \, n \cdot v|_{\partial\Omega} = 0\}, \\ &G^r = \{v = \nabla \phi; \, \phi, \nabla \phi \in L^r(\Omega)\},\end{aligned} \tag{4.11}$$

provided $1<r<\infty$. Let P be the projection to L^r_σ associated with the decomposition. The Stokes operator $A=-P\Delta$ is defined on

$$D(A)=\{v\in L^r_\sigma;\ \nabla^2 v\in L^r(\Omega),\ v|_{\partial\Omega}=0\}.$$

By a priori estimates [17], we see

$$\|\nabla^2 u\|_r \leqq C\|Au\|_r, \qquad u\in D(A), \tag{4.12}$$

with C independent of u. The Stokes operator generates an analytic semigroup e^{-tA} on L^r_σ (cf. [8], [17]), and $w=e^{-tA}w_0$ with $w_0\in L^r_\sigma$ solves the Stokes system

$$\partial_t w-\Delta w+\nabla p=0, \qquad \operatorname{div} w=0 \quad \text{in } \Omega\times(0,T),$$

with

$$w|_{\partial\Omega}=0 \quad \text{and} \quad w(x,0)=w_0\in L^r_\sigma.$$

Since e^{-tA} is an analytic semigroup and (4.12) holds, we observe that $w=e^{-tA}w_0$ is smooth for $t>0$. For $v\in E_r$ we set

$$v_\varepsilon=e^{\varepsilon A}v(\cdot,t), \qquad \varepsilon<0$$

which is smooth in x for almost everywhere t. We observe

$$v_\varepsilon\to v \quad \text{in } E_r. \tag{4.13}$$

Indeed, by (4.12) we see

$$\begin{aligned}\int_0^T \|\nabla^2(e^{\varepsilon A}v-v)\|_r\,ds &\leqq C\int_0^T \|A(e^{\varepsilon A}v-v)\|_r\,ds\\ &= C\int_0^T \|(e^{\varepsilon A}-I)Av\|_r\,ds.\end{aligned}$$

Since $\|(e^{\varepsilon A}-I)w\|_r\to 0$ for $w\in L^r_\sigma$ as $\varepsilon\to 0$, and since $\|(e^{\varepsilon A}-I)Av\|_r\leqq 2\|Av\|_r$, applying Lebesgue's convergence theorem yields

$$\int_0^T \|\nabla^2(e^{\varepsilon A}v-v)\|_r\,ds\to 0 \quad \text{as } \varepsilon\to 0.$$

This yields (4.13) by using the Sobolev inequality. Since the function v_ε is approximated by a $C^\infty(\bar{Q}_T)$ function in E_r by smoothing the time direction, we conclude that Q is dense in E_r. □

PROPOSITION 4.4. *Suppose that v_i $(i=1,2)$ satisfies* (4.2), (4.3) *with $v_i\in E_r$, $r>N$, and that $\eta=\eta_i$ solves* (4.1), (4.4) *with $v=v_i$ and satisfies* (4.10) *for $\eta_i|_{t=0}=\eta_{i,0}\in \mathrm{BV}(\Omega)\cap L^\infty(\Omega)$. Then,*

$$\|\eta_1-\eta_2\|_1(t)\leqq L\int_0^T \|v_1-v_2\|_\infty(s)\,ds+\|\eta_{1,0}-\eta_{2,0}\|_1, \tag{4.14}$$

where L is an upper bound of

$$\max_{i=1,2}\left\{C\|\nabla\eta_{i,0}\|_1\exp\int_0^T \|\nabla v_i\|_\infty\,ds\right\},$$

with C depending only on N.

Proof. We first assume that $v_i\in C^\infty(\bar{Q}_T)$ and $\eta_{i,0}\in C^\infty(\bar{\Omega})$. By a representation of η_i, we see $\eta_i\in C^\infty(\bar{Q}_T)$. The difference $w=\eta_1-\eta_2$ solves

$$\partial_t w+(v_1\cdot\nabla)w=-(v_1-v_2)\nabla\eta_2, \qquad w|_{t=0}=\eta_{1,0}-\eta_{2,0}. \tag{4.15}$$

Since $(v_2-v_1)\nabla\eta_2 \in C^\infty(\bar{Q}_T)$, Lemma 4.1 shows that

$$\phi(x,t)=\int_0^t [(v_2-v_1)\nabla\eta_2](y_1(t,s,x),s)\,ds+[\eta_{1,0}-\eta_{2,0}](y_1(t,0,x))$$

solves (4.15) with $\phi = w$, where y_1 is a characteristic curve corresponding to v_1, viz., y_1 is defined by (4.5) with $v = v_1$. By the uniqueness of the solution in Lemma 4.2, we see ϕ agrees with our w. In a similar way, to obtain (4.8) we have

$$\begin{aligned}\|w\|_1(t) &\leqq \int_0^t \|v_2-v_1\|_\infty(s)\|\nabla\eta_2\|_1(s)\,ds+\|\eta_{1,0}-\eta_{2,0}\|_1\\ &\leqq L\int_0^t \|v_2-v_1\|_\infty(s)\,ds+\|\eta_{1,0}-\eta_{2,0}\|_1,\end{aligned}$$

since (4.7) holds.

We approximate $v \in E_r$ and η_0. By Lemma 4.3 there is a sequence $\{v^m\} \in C^\infty(\bar{Q}_T) \cap E_r$ such that

$$v^m \to v \quad \text{in } E_r \quad \text{as } m \to \infty. \tag{4.16}$$

For $\eta_0 \in \mathrm{BV}(\Omega) \cap L^\infty(\Omega)$, there is a sequence $\{\eta_0^m\} \in C^\infty(\bar{\Omega})$ such that

$$\eta_0^m \to \eta \qquad \text{in } L^1(\Omega) \quad \text{as } m \to \infty, \tag{4.17}$$

$$\|\eta_0^m\|_\infty \leqq \|\eta_0\|_\infty, \qquad \|\nabla\eta_0^m\|_1 \leq C\|\nabla\eta_0\|_1. \tag{4.18}$$

If $\Omega = \mathbf{R}^N$, we find approximate sequences $\{\eta_0^m\}$ of η_0 just by mollifying η_0; see [10]. If Ω has the boundary, we reduce the problem when Ω is a half-space by a localization. Since the BV function has the trace on the boundary (see [10]), we can extend the BV function outside the half-space without increasing $\|\nabla\eta\|_1$ and $\|\eta\|_\infty$. The problem is now reduced to the case when $\Omega = \mathbf{R}^N$.

Let η^m solve

$$\partial_t\eta^m + (v^m\cdot\nabla)\eta^m = 0, \qquad \eta^m|_{t=0} = \eta_0^m. \tag{4.19}$$

We observe that η^m actually approximates the solution η of

$$\partial_t\eta + (v\cdot\nabla)\eta = 0, \qquad \eta|_{t=0} = \eta_0. \tag{4.20}$$

Applying (4.14) for smooth data, we see $\{\eta^m\}$ is a Cauchy sequence in $C([0,T]; L^1(\Omega))$ since (4.16)–(4.18) holds. Let η denote its limit. Since $v^m \to v$ in $L^1(0,T; L^\infty(\Omega))$, we see $v^m\eta^m$ converges to $v\eta$ in $L^1(Q_T)$. This implies that η solves

$$\partial_t\eta + \operatorname{div}(v\eta) = 0, \qquad \eta|_{t=0} = \eta_0. \tag{4.21}$$

Since $r > N$, we have

$$\int_0^T [\|v\|_\infty + \|\nabla v\|_\infty]\,ds \leqq C|||v|||_r$$

by the Sobolev inequality. By Lemma 4.2 (ii), (iii), we see $\{\eta^m\}$ and $\{\nabla\eta^m\}$ are bounded in $L^\infty(Q_T)$ and $L^\infty(0,T; M(\Omega))$, respectively, since (4.16)–(4.18) hold. Also $\{\partial_t\eta^m\}$ is bounded in $L^1(0,T; M(\Omega))$ by (4.19). We thus observe that the limit η satisfies (4.10). This implies that η solves (4.20) since η solves (4.21) and $\operatorname{div} v = 0$. By uniqueness (Lemma 4.2) we conclude our η is the only solution of (4.20) and

$$\eta^m \to \eta \quad \text{in } C([0,T]; L^1(\Omega)). \tag{4.22}$$

It remains to prove (4.14) for $v_i \in E_r$ and $\eta_{i,0} \in \mathrm{BV}(\Omega) \cap L^\infty(\Omega)$. We approximate each v_i, $\eta_{i,0}$ by v_i^m, $\eta_{i,0}^m$ defined above, and pass the limit. Applying (4.16), (4.17), and (4.22) now yields (4.14) for general initial data. □

5. A priori estimates for parabolic systems with discontinuous coefficients. We consider a parabolic system

$$\partial_t b = -\mathrm{rot}^*\eta \ \mathrm{rot}\, b - \beta \ \mathrm{grad}\ \mathrm{div}\, b + \mathrm{rot}^* f \quad \text{in } Q_T = \Omega \times (0, T),$$

with the boundary conditions

$$n\Lambda^* \ \mathrm{rot}\, b = 0, \qquad n \cdot b = 0 \quad \text{on } \partial\Omega \times (0, T),$$

and initial condition

$$b|_{t=0} = 0.$$

Here f and η are given functions and $0 < \alpha \leqq \eta \leqq \beta$ with constants α and β. Our goal is to derive a space-time L^r estimate for ∇b, when f is in $L^r(Q_T)$ and satisfies a compatibility condition $n\Lambda^* f = 0$ on $\partial\Omega$, even if η is discontinuous. When the boundary condition is of Dirichlet type, Campanato [4] establishes such a result for r very close to 2 by using a perturbation argument. We also apply a perturbation argument; however, we need more. We approximate η by smooth functions so that the trace $n\Lambda^* \ \mathrm{rot}\, b$ is defined at least for smooth f. We first consider the initial value problem for the heat equation

$$\partial_t U - \beta \Delta U = \mathrm{rot}^* f \quad \text{in } Q_T, \tag{5.1}$$

$$n\Lambda^* \ \mathrm{rot}\, U = 0, \quad n \cdot U = 0 \quad \text{on } \partial\Omega \times (0, T), \tag{5.2}$$

$$U|_{t=0} = 0, \tag{5.3}$$

where Ω is a smoothly bounded domain in $\mathbf{R}^N$ ($N = 2, 3$). As is well known (see [12]), there is a unique solution $U \in W_2^{2,1}(Q_T)$ of (5.1)-(5.3) if f satisfies

$$f, \nabla f \in L^2(Q_T) \quad \text{for some } T > 0. \tag{5.4}$$

If f is smooth in $\bar{\Omega} \times (0, T)$, then so is U (cf. [6], [12]). Throughout this section, we will assume that f satisfies (5.4) to understand (5.2) by usual identities of traces of functions without using a weak formulation of (5.1)-(5.3).

LEMMA 5.1. *Suppose that U solves* (5.1)-(5.3). *Then we have*

$$[\nabla U]_{r,T} \leqq A_r \beta^{-1} [f]_{r,T}, \tag{5.5}$$

where

$$[f]_{r,T}^r = \int_0^T \|f\|_r^r(s)\, ds,$$

and the constant A_r is independent of f, T, and β. Suppose furthermore that $n\Lambda^* f = 0$ on $\partial\Omega$. Then we may take A_r such that $A_r \to 1$ as $r \to 2$.

Proof. Except for the last assertion, these are essentially known as a priori estimates. Such estimates are found in [14] for single elliptic equations with Neumann boundary condition. However, since our problem is a system, we do not find a suitable version in the literature, so we give here an outline of the proof. We may assume that $\beta = 1$ by setting $\bar{U}(x, t) = U(x, t/\beta)$. We observe that (5.5) holds when Ω is $\mathbf{R}^N$ or a half-space $\mathbf{R}_+^N$ by applying the Calderon-Zygmund inequality. For general Ω we appeal to a usual cutoff procedure and changes of coordinates near the boundary so that the boundary becomes flat in the new coordinates. The operator may have variable

coefficients because of the coordinate changes. Since $\partial\Omega$ is smooth, so are the coefficients. Applying a standard method of freezing coefficients, we deduce from (5.5) for $\mathbf{R}^N$ and $\mathbf{R}^N_+$ that

$$[\nabla U]_{r,T} \leqq C_r([f]_{r,T}+[U]_{r,T}). \tag{5.6}$$

The estimate (5.5) follows from (5.6) and the uniqueness of the solutions of (5.1)-(5.3). Indeed, we may assume $T \geqq 1$. Suppose that (5.5) were false for $T \geqq 1$. Then there would exist a sequence of functions $\{f_l\}$ and $\{U_l\}$ (the corresponding solutions of (5.1)-(5.3)) such that

$$[\nabla U_l]_{r,T(l)} = 1, \qquad [f_l]_{r,T(l)} < \frac{1}{l} \tag{5.7}$$

with some sequence $\{T(l)\}$ converging to $T \geqq 1$, which may be ∞. Since Ω is bounded and U_l satisfies (5.2), the Poincaré inequality and (5.7) imply that

$$U_l \to V, \qquad \nabla U_l \to \nabla V \quad \text{weakly in } L^r(Q_T) \tag{5.8}$$

with some V by taking a subsequence. A standard argument for parabolic equations shows that

$$U_l \to V \quad \text{strongly in } L^r(Q_T).$$

We sketch the argument for completeness. We multiply $|U_l|^{r-2}U_l$ with (5.1) for $U = U_l$, and integrate by parts to get an energy estimate which yields

$$\sup_{l \geqq 1} \sup_{0 \leqq t \leqq T} \|U_l\|_r(t) < \infty. \tag{5.9}$$

Since U_l solves (5.1)-(5.3) with (5.8), we also observe that

$$\lim_{l \to \infty} \int_\Omega U_l(x, t)\phi(x)\, dx$$

exists for every $\phi \in C_0^1(\Omega)$. By (5.8) and (5.9), this yields

$$U_l(\cdot, t) \to V(\cdot, t) \quad \text{weakly in } L^r(\Omega) \quad \text{for a.e. } t.$$

We may regard this convergence as the strong one in $L^r(\Omega)$ by (5.7), (5.9), and Rellich's theorem. By (5.9) we apply Lebesgue's convergence theorem to get the strong convergence $U_l \to V$ in $L^r(Q_T)$. By (5.6) and (5.7) we see

$$C_r[V]_{r,T} \geqq 1, \tag{5.10}$$

since $U_l \to V$ strongly in $L^r(Q_T)$. Since U_l solves (5.1)-(5.3), V solves a weak form of (5.1)-(5.3) with $f = 0$; the condition $n \cdot V = 0$ follows from $n \cdot U_l = 0$ on $\partial\Omega$. We now apply the following uniqueness of solutions.

LEMMA 5.2. *Suppose that $\eta \in L^\infty(Q_T)$ and that $\alpha \leqq \eta \leqq \beta$ with two positive constants α and β. Suppose that V with $\nabla V \in L^r(Q_T)$ satisfies*

$$\int_0^T \int_\Omega V\partial_t\phi\, dx\, dt + \int_0^T \int_\Omega [\eta\ (\mathrm{rot}\ V \cdot \mathrm{rot}\ \phi) + \beta(\mathrm{div}\ V \cdot \mathrm{div}\ \phi)]\, dx\, dt = 0, \tag{5.11}$$

for every $\phi \in C_0^1(\bar{\Omega} \times [0, T))$ with $n \cdot \phi = 0$ on $\partial\Omega$. Suppose that $n \cdot V = 0$ on $\partial\Omega$. Then $V = 0$ on Q_T.

Applying Lemma 5.2 with $\eta = \alpha = \beta = 1$, we conclude that $V = 0$ which contradicts (5.10). We thus obtain (5.5) for $T \geqq 1$.

It remains to prove that $A_r \to 1$ as $r \to 2$ for a suitable choice of A_r provided that $n\Lambda^* f = 0$ on $\partial\Omega$. We multiply U with (5.1) and integrate over Ω. Integrating by parts with (5.2) yields

$$\frac{d}{2dt}\|U\|_2^2(t) + \|\nabla U\|_2^2(t) = \int_\Omega \operatorname{rot}^* f \cdot U\, dx \quad \text{a.e. } t.$$

Since

$$\int_\Omega \operatorname{rot}^* f \cdot U\, dx = \int_\Omega f \cdot \operatorname{rot} U\, dx + \int_{\partial\Omega} (n\Lambda^* f) U\, dS,$$

and $n\Lambda^* f = 0$ on $\partial\Omega$, we now have

$$\frac{d}{2dt}\|U\|_2^2(t) + \|\nabla U\|_2^2(t) \leqq \|f\|_2 \|\nabla U\|_2 \quad \text{a.e. } t,$$

where dS denotes the surface element. Integrating over $(0, t)$ and noting (5.3) yields (5.5) with $A_r = 1$. Interpolating (5.5) for $p = p_0 > 2$ or $p_0 < 2$ and $p = 2$ by the Riesz-Thorin theorem (see, e.g., [16]), we obtain the desired choice of A_r (cf. Lemma 1.II of Campanato [4]).

Proof of Lemma 5.2. We only give a proof when $r \geqq 2$ because we do not use the case $r < 2$. Since $\nabla V \in L^2(Q_T)$ and $C^1(\bar{\Omega})$ is dense in $H^1(\Omega)$, we may plug $\phi \in C_0^1([0, T); H^1(\Omega))$ in (5.11) with $n \cdot \phi = 0$ on $\partial\Omega$. We extend $V = 0$ for $t < 0$ and $t > T$, and consider $V_\delta = \rho_\delta * \rho_\delta * V$ where ρ_δ is a Friedrich's mollifier on $\mathbf{R}$ with $\rho_\delta(x) = \rho_\delta(-x)$ and $*$ denotes the convolution on $\mathbf{R}$. Let θ_δ be a cutoff function of $(-\infty, T - \delta)$ such that $\theta_\delta = 1$ on $(-\infty, T - \delta)$, $\theta_\delta = 0$ on (T, ∞), $0 \leqq \theta_\delta \leqq 1$, and $0 \leqq -\partial_t \theta_\delta \leqq 2/\delta$. We plug $\phi = \theta_\delta V_\delta$ in (5.11) and obtain

$$\frac{1}{2}\int_{-\infty}^{T} \theta_\delta(s) \frac{d}{ds}\|V^\delta\|_2(s)\, ds + \int_0^T \theta_\delta \int_\Omega [\eta |\operatorname{rot} V^\delta|^2 + \beta |\operatorname{div} V^\delta|^2]\, dx\, ds = 0,$$

where $V^\delta = \rho_\delta * V$. Since

$$\int_\Omega [|\operatorname{rot} V^\delta|^2 + |\operatorname{div} V^\delta|^2]\, dx = \int_\Omega |\nabla V^\delta|^2\, dx,$$

letting $\delta \to 0$ implies $V = 0$ (cf. Temam [18, p. 311]). □

PROPOSITION 5.3. *Suppose that* $u \in W_2^{2,1}(Q_T)$ *solves*

$$\partial_t u = -\operatorname{rot}^* \eta \operatorname{rot} u - \beta \operatorname{grad} \operatorname{div} u + \operatorname{rot}^* f \quad \text{in } Q_T, \tag{5.12}$$

$$n\Lambda^* \operatorname{rot} u = 0, \qquad n \cdot u = 0 \quad \text{on } \partial\Omega \times (0, T), \tag{5.13}$$

$$u|_{t=0} = 0 \quad \text{on } \Omega. \tag{5.14}$$

Suppose that $\eta \in C^\infty(\bar{Q}_T)$ *satisfies* $\alpha \leqq \eta \leqq \beta$ *in* Q_T, *where* α *and* β *are two positive constants, and that* f *satisfies* (5.4). *Then there is a function* $p_0(\sigma) > 2$ *defined on* $\sigma > 0$ *such that*

$$[\nabla u]_{r,T} \leqq K[f]_{r,T} \quad \text{for } 2 \leqq r < p_0(\alpha/\beta), \tag{5.15}$$

$$p_0(\sigma) \to \infty \quad \text{as } \sigma \to 1, \tag{5.16}$$

with $K = [1 - A_r(1 - \alpha/\beta)]^{-1} A_r \beta^{-1}$, *provided that*

$$n\Lambda^* f = 0 \quad \text{on } \partial\Omega.$$

Here A_r *is defined in Lemma* 5.1.

Proof. We rewrite (5.12) and obtain

$$\partial_t u - \beta \Delta u = \operatorname{rot}^* F,$$

where

$$F = f - \eta \operatorname{rot} u + \beta \operatorname{rot} u.$$

By (5.13) we see F fulfills all assumptions for f in Lemma 5.1, since $u \in W_2^{2,1}(Q_T)$. Applying Lemma 5.1 yields

$$\begin{aligned}[\nabla u]_{r,T} &\leqq A_r \beta^{-1}([f]_{r,T} + \sup_x |\beta - \eta|[\nabla u]_{r,T}) \\ &\leqq A_r \beta^{-1}[f]_{r,T} + A_r(1 - \alpha/\beta)[\nabla u]_{r,T}.\end{aligned}$$

If $A_r(1-\alpha/\beta) < 1$, then this inequality yields (5.15). Since $A_r \to 1$ as $r \to 2$ and $A_2 = 1$, we now conclude that there is $p_0 > 2$ satisfying (5.15) and (5.16). □

We extend these L^r estimates to the case when η is discontinuous. Since the solution of (5.12)-(5.14) is no longer expected to be in $W_2^{2,1}(Q_T)$, we need a weak formulation similar to (3.9). We say u is a weak solution of (5.12)-(5.14) if u satisfies

$$\begin{aligned}&\int_0^T \int_\Omega u \cdot \partial_t \phi \, dx \, dt \\ &\quad = -\int_0^T \bar{a}(t, u, \phi) \, dt + \int_0^T \int_\Omega \operatorname{rot}^* f \cdot \phi \, dx \, dt \quad \text{with } n \cdot u = 0 \quad \text{on } \partial\Omega\end{aligned} \tag{5.17}$$

for every $\phi(x, t) \in C_0^1(\bar{\Omega} \times [0, T))$, where

$$\bar{a}(t, u, \phi) = \int_\Omega \eta(x, t) \operatorname{rot} u \cdot \operatorname{rot} \phi \, dx + \int_\Omega \beta (\operatorname{div} u)(\operatorname{div} \phi) \, dx. \tag{5.18}$$

PROPOSITION 5.4. *Suppose that u with $\nabla u \in L^2(Q_T)$ solves* (5.17) *with f satisfying* (5.4), *and that $\eta \in L^\infty(Q_T)$ satisfies $0 < \alpha \leqq \eta \leqq \beta$. Then, there is a function $p_0(\sigma) > 2$ defined for $\sigma > 0$ such that* (5.15) *holds for $2 \leqq r < p_0(\alpha/\beta)$ and that* (5.16) *holds provided that $n\Lambda^* f = 0$ on $\partial\Omega$.*

Proof. There is $\eta_\delta \in C^\infty(\bar{Q}_T)$ such that $\eta_\delta \to \eta$ in $L^q(Q_T)$, $q > 1$ strongly as $\delta \to 0$ and $\alpha \leqq \eta_\delta \leqq \beta$. By a standard parabolic theory, there is a unique solution u_δ which solves (5.12)-(5.14) with setting $\eta = \eta_\delta$; see [6], [12]. Similarly to the last part of the proof of Lemma 5.1, we multiply u_δ with (5.12), and integrate over space-time. Integrating by parts yields energy inequality

$$\|u_\delta\|_2^2(t) + \alpha \int_0^t \int_\Omega |\nabla u_\delta|^2 \, dx \, ds \leqq \int_0^t \int_\Omega |f|^2 \, dx \, dt.$$

As in the proof of Lemma 5.1, this estimate together with a standard argument for parabolic equations implies that there is some $\bar{u}$ such that

$$\begin{aligned}&u_\delta \to \bar{u} \quad \text{strongly in } L^2(Q_T), \\ &\nabla u_\delta \to \nabla \bar{u} \quad \text{weakly in } L^2(Q_T)\end{aligned}$$

by passing to subsequences. Since $\eta_\delta \to \eta$ in $L^q(Q_T)$ for every $q > 1$, we conclude that $\bar{u}$ solves (5.17); the condition $n \cdot \bar{u} = 0$ on $\partial\Omega$ follows from $n \cdot u_\delta = 0$. The uniqueness of solution (Lemma 5.2) implies that $\bar{u} = u$. We thus conclude that (5.15) and (5.16) still hold for $u = \bar{u}$, since the norm $[\nabla u]_{r,T}$ is lower semicontinuous under the weak topology in $L^2(Q_T)$. □

Remark. The existence of solutions u of (5.17) follows from a generalized Lax-Milgram lemma (cf. Lions and Magenes [13, Chap. 3, Thm. 4.1] with a remark on the Neumann problem [13, Chap. 3, 4.7.2]. However, for $u, \nabla u \in L^2(Q_T)$, the meaning of $n\Lambda^*$ rot u is unclear, so it seems difficult to apply a perturbation argument of the proof of Proposition 5.3 directly to Proposition 5.4 without approximations.

6. Proof of the Main Existence Theorem. We will prove that our F in § 3 has a fixed point in a suitable class of functions by using a priori estimates we obtained in §§ 4 and 5. We begin with existence and uniqueness of solutions of the initial value problem for (2.9), (2.10), (2.12) assuming that η is a given function.

PROPOSITION 6.1. *Suppose that B is smooth on $\bar{\Omega}$ and satisfies* (2.4). *Suppose that $\eta \in L^\infty(Q_T)$ and $\alpha \leqq \eta \leqq \beta$ with positive constants α and β. Then, for initial data $v|_{t=0} = v_0$, $b|_{t=0} = b_0 \in L^2_\sigma(\Omega)$ there is a unique function $(b(x, t), v(x, t))$ that solves* (3.9), (2.10), (3.10), *and* (3.11) *with some $p(x, t)$ such that*

$$b \in C([0, T]; L^2_\sigma(\Omega)) \cap L^2(0, T; W), \qquad W = H^1(\Omega) \cap L^2_\sigma(\Omega), \tag{6.1}$$

$$v \in C([0, T]; L^2_\sigma(\Omega)) \cap L^2(0, T; W_0), \qquad W_0 = \{v \in W;\ v = 0 \text{ on } \partial\Omega\}. \tag{6.2}$$

Moreover, (b, v) satisfies the energy equality

$$\|b\|_2^2(t) + \|v\|_2^2(t) + 2\int_0^t \int_\Omega \eta(\text{rot } b)^2\, dx\, dt + 2\int_0^t \int_\Omega (\nabla v)^2\, dx\, dt = \|b_0\|_2^2 + \|v_0\|_2^2, \qquad 0 \leqq t \leqq T. \tag{6.3}$$

Proof. We set

$$A(t, u, u') = a(t, b, b') - \int_\Omega (\text{rot}^*(v\Lambda B) \cdot b')\, dx + \int_\Omega \nabla v \cdot \nabla v'\, dx + \int_\Omega (B\Lambda^* \text{ rot } b) \cdot v'\, dx$$

for $u = (b, v)$, $u' = (b', v')$, where a is defined by (3.8). We set $V = W \times W_0$ and observe that A is coercive on V, since

$$A(t, u, u) = a(t, b, b) + \int_\Omega |\nabla v|^2\, dx \geqq \alpha \|\text{rot } b\|_2^2 + \|\nabla v\|_2^2,$$

and the right-hand side is a norm of V. We also observe that A is continuous on V and $t \mapsto A(t, u, u')$ is measurable in t for $u, u' \in V$. Applying a generalized Lax-Milgram theorem (Lions and Magenes [13, Chap. 3, Thm. 4.1, Remark 4.3]), we conclude that there is a unique solution $u = (b, v)$ such that

$$u \in L^2(0, T; V) \cap C([0, T]; H), \qquad H = L^2_\sigma(\Omega) \times L^2_\sigma(\Omega), \tag{6.4}$$

and that at least satisfies

$$\int_0^T A(t, u(t), \Psi(t))\, dt - \int_0^T \int_\Omega (u(x, t) \cdot \partial_t \Psi)\, dx\, dt = \int_\Omega u(x, 0) \cdot \Psi(x, 0)\, dx \tag{6.5}$$

for every $\Psi=(\phi, \psi)$, where $u(t)=u(\cdot, t)$. From this we deduce (2.10) and (3.9) at least for test functions ϕ satisfying $\phi(\cdot, t) \in L^2_\sigma(\Omega)$. For arbitrary $\bar{\phi} \in C^1_0([0, T); H^1(\Omega))$, we have a Helmholtz decomposition such that

$$\bar{\phi}=\phi+\nabla p, \qquad \phi(\cdot, t) \in L_\sigma(\Omega) \cap H^1(\Omega) \tag{6.6}$$

and

$$\|\phi\|_{H^1}(t), \|\nabla p\|_{H^1}(t) \leqq C\|\bar{\phi}\|_{H^1}(t), \tag{6.7}$$

with C independent of t and $\bar{\phi}$. In other words, the projection P associated with Helmholtz decomposition (4.11) is continuous in H^1 topology (see [9, Lemma 3.3]). We observe that

$$\int_0^T \int_\Omega b \cdot \partial_t(\nabla p)\, dx\, dt=0, \qquad \int_\Omega b_0 \cdot \nabla p(x, 0)\, dx=0,$$

since $b(\cdot, t) \in L^2_\sigma(\Omega)$. Since $v(t) \in W_0$ implies $v=0$ on $\partial\Omega$ and $\operatorname{rot}(\nabla p)=0$, we have

$$\int_\Omega \operatorname{rot}^*(v \Lambda B) \cdot \nabla p\, dx=\int_\Omega (v \Lambda B) \cdot \operatorname{rot}(\nabla p)\, dx+\int_{\partial\Omega} n \Lambda^*(v \Lambda B) \cdot \nabla p\, dS=0$$

and

$$\int_0^T a(t, b, \nabla p)=0.$$

These calculations are justified by (6.7). We conclude that (3.9) holds for every $(b, v) \in L^2(0, T; V)$ if $\phi=\nabla p$. We thus see (3.9) holds for every $\phi \in C^1_0([0, T); H^1(\Omega))$ because of (6.6). The regularity (6.1), (6.2) follows directly from (6.4). Condition (3.10) follows from $v(t) \in W_0$.

It remains to prove (6.3). This is obtained by plugging $\Psi=u$ in (6.5). To justify this calculation we should approximate u by test functions. The idea is similar to the proof of Lemma 5.2 and it is standard so the proof is omitted. □

We next study regularity of (b, v). The regularity of v in (6.2) is not enough to solve the transport equation uniquely in BV spaces. We apply a priori estimates for (3.9) and derive a suitable regularity of v through (2.10).

PROPOSITION 6.2. *Let (b, v) be a solution of* (3.9), (2.10), (3.10) *given in Proposition* 6.1.

(i) *Suppose that the initial data b_0 satisfies* (3.3) *with $q>1$. Then*

$$[\nabla b]_{r,T} \leqq K(C_B[v]_{r,T}+\beta T^\sigma \|\operatorname{rot} b_0\|_r), \qquad \sigma=1/r, \tag{6.8}$$

for $2 \leqq r<\min(q, p_0(\alpha/\beta))$, where p_0 and K are defined in Proposition 5.3 *and $C_B=\sup|B|$.*

(ii) *Suppose that the initial data v_0 satisfies* (3.2), *i.e., $v_0 \in W^2_q(\Omega)$. Then*

$$[\nabla^2 v]_{r,T} \leqq C([\nabla b]_{r,T}+\|v_0\|_{W^2_r(\Omega)}), \qquad 1<r \leqq q, \tag{6.9}$$

with some C depending only on r, T, N, and Ω.

(iii) *Suppose* (3.2) *and* (3.3) *with $q>N$. Then we have*

$$[\nabla^2 v]_{r,T} \leqq Z, \tag{6.10}$$

with some r, $N < r \leqq q$, and a constant Z depending on data as

$$Z = Z(\beta/\alpha, \beta, T, \|v_0\|_{W^2_r(\Omega)}, \|\operatorname{rot} b_0\|_r, r, B), \tag{6.11}$$

provided that either

(a) $N = 2$, *or*

(b) $N = 3$, *and β/α is sufficiently close to* 1

holds.

Proof. (i) Since $\operatorname{div} b = 0$, we may add $-\beta \operatorname{grad} \operatorname{div} b$ to both sides of (2.9). We set $\tilde{b} = b - b_0$ and rewrite (2.9) to get (5.12) (or precisely, its weak form (5.17)) with $u = \tilde{b}$ and $f = v\Lambda B - \eta \operatorname{rot} b_0$. Since (6.2) and (3.3) imply that f satisfies (5.4), we now apply Proposition 5.4 to obtain (6.8).

(ii) This follows directly from a priori L^r estimates for the Stokes system (2.10) due to Solonnikov [17], where $-B\Lambda^* \operatorname{rot} b$ should be regarded as an external force.

(iii) Since (6.3) yields

$$\sup_{t \geqq 0} \|v\|_2^2, 2[\nabla v]_{2,T}^2 \leqq \|v_0\|_2^2 + \|b_0\|_2^2 = \omega,$$

we first observe that

$$[v]_{4,T} \leqq C\omega^{1/2} \quad \text{for } N = 2, \tag{6.12a}$$

$$[v]_{10/3,T} \leqq C\omega^{1/2} \quad \text{for } N = 3, \tag{6.12b}$$

with C independent of b, v_0, and B. Applying the Gagliardo-Nirenberg inequality yields (6.12a, b) (cf. [3, p. 781]).

Suppose that $N = 2$. Then combining (6.12a), (6.8), (6.9) yields (6.10) for all r such that $2 \leqq r < \min(q, p_0(\alpha/\beta), 4)$. When $N = 3$, by (5.16) we can take β/α close to 1 so that $p_0(\beta/\alpha) > N = 3$. Since $10/3 > 3$, combining (6.12b), (6.8), (6.9) now yields (6.10) for all r such that $3 < r < \min(q, p_0(\alpha/\beta), 10/3)$. □

For a given $0 < \alpha \leqq \beta$ we consider a class of resistivity functions:

$$S(L) = \Big\{ \eta(x, t) \in \mathrm{BV}(Q_T);\ \alpha \leqq \eta \leqq \beta \text{ a.e. } (x, t) \in Q_T, \tag{6.13}$$
$$\sup_{0 \leqq t \leqq T} \|\nabla \eta\|_1(t) \leqq L, \int_0^T \|\partial_t \eta\|_1 \, ds \leqq L \Big\}.$$

For $\eta \in S(L)$ we construct a solution (b, v) of (3.9), (2.10), (3.10) by Proposition 6.1; we will write $(b, v) = (G_1(\eta), G_2(\eta))$. If we assume (3.2) and (3.3) with $q > N$, $v = G_2(\eta)$ satisfies (6.10) provided that either (a) or (b) in Proposition 6.2 holds. Here and hereafter we will always assume that (a) or (b). By the Sobolev inequality and (6.12a, b) we observe that v satisfies (4.3) and moreover

$$\int_0^T (\|v\|_\infty + \|\nabla v\|_\infty) \, ds \leqq CZ, \tag{6.14}$$

with C depending only on Ω, N, and T. Applying Lemma 4.2 now implies that, for $v = G_2(\eta)$, the transport equation (2.12) has a unique solution $\eta = E(v)$ satisfying (3.4), if the initial data η_0 satisfies (3.1). We will write

$$F(\eta) = E(G_2(\eta)).$$

PROPOSITION 6.3. *Suppose that* (3.1)-(3.3) *holds for $q > N$. If L is sufficiently large, then F maps $S(L)$ into itself.*

Proof. By Lemma 4.2(ii) and (6.14) we see

$$\|\nabla\zeta\|_1(t) \leqq A\|\nabla\eta_0\|_1 e^{aCZ} = Z', \qquad 0 \leqq t \leqq T,$$

$$\int_0^T \|\partial_t\zeta\|_1(s)\,ds \leqq \int_0^T \|v\|_\infty \|\nabla\zeta\|_1(s)\,ds \leqq CZ \sup_{0\leqq t\leqq T} \|\nabla\zeta\|_1$$

$$\leqq CZ\|\nabla\eta_0\|_1 e^{CZ} = CZZ',$$

with $\zeta = F(\eta)$. Since Z' and CZZ' are independent of L, we take L large such that $Z', CZZ' \leqq L$. Since $\alpha \leqq \eta_0 \leqq \beta$ implies $\alpha \leqq \zeta \leqq \beta$ by Lemma 4.2(iii), we conclude that F maps $S(L)$ into itself for L defined above. □

PROPOSITION 6.4. *The set $S(L)$ is convex and compact in $L^\rho(Q_T)$, $1 \leqq \rho < \infty$.*

Proof. Clearly $S(L)$ is convex. Since $S(L)$ is bounded in $\mathrm{BV}(Q_T)$, it is relatively compact in $L^h(Q_T)$, $1 \leqq h < (N+1)/N$ by Rellich's theorem [10]. Since $S(L)$ is bounded in $L^\infty(Q_T)$, this means that $S(L)$ is relatively compact in every $L^\rho(Q_T)$, $1 \leqq \rho < \infty$. Since $\sup \|\nabla\eta\|_1$ and $\int \|\partial_t\eta\|_1\,ds$ are lower semicontinuous in L^ρ topology, we conclude that $S(L)$ is compact in $L^\rho(Q_T)$. □

PROPOSITION 6.5. *Suppose that* (3.1)–(3.3) *holds for $q > N$. Then F is continuous under the topology of $L^\rho(Q_T)$, $1 \leqq \rho < \infty$.*

Proof. For $\eta_1, \eta_2 \in S(L)$ we set $(b_i, v_i) = (G_1(\eta_i), G_2(\eta_i))$, $\zeta_i = F(\eta_i)$ $(i = 1, 2)$, $\bar\eta = \eta_1 - \eta_2$, and $\bar\zeta = \zeta_1 - \zeta_2$. We will prove that

$$[\bar\zeta]_{1,T} \leqq c[\bar\eta]_{p,T} \tag{6.15}$$

with some p, $1 \leqq p < \infty$, and c independent of η_1 and η_2. We first derive the estimate for $\bar b = b_1 - b_2$. For technical reasons we approximate b_i by a solution of (2.9) with mollified coefficients in the term of the second order. As in the proof of Proposition 5.4, we take $\eta_{i,\delta} \in C^\infty(\bar Q_T)$ such that $\eta_{i,\delta} \to \eta_i$ in $L^h(Q_T)$ strongly for $h > 1$ as $\delta \to 0$ and $\alpha \leqq \eta_{i,\delta} \leqq \beta$. Let $b_{i,\delta}$ denote the solution of (5.12), (5.13) with $b_{i,\delta}|_{t=0} = b_0$, $\eta = \eta_{i,\delta}$, $f = v_i \Lambda B$. As in the proof of Proposition 5.4 we see

$$\begin{aligned} &b_{i,\delta} \to b_i \quad \text{strongly in } L^2(Q_T), \\ &\nabla b_{i,\delta} \to \nabla b_i \quad \text{weakly in } L^2(Q_T), \end{aligned} \tag{6.16}$$

since b_i uniquely solves (5.17) by Lemma 5.2. Since the difference $\bar b_\delta = b_{1,\delta} - b_{2,\delta}$ solves

$$\begin{aligned} &\partial_t \bar b_\delta = -\mathrm{rot}^*\,\eta_{1,\delta}\,\mathrm{rot}\,\bar b_\delta - \mathrm{rot}^*\,\bar\eta_\delta\,\mathrm{rot}\,b_{2,\delta} + \mathrm{rot}^*(\bar v \Lambda B), \\ &\mathrm{div}\,\bar b_\delta = 0, \\ &n\Lambda^*\,\mathrm{rot}\,\bar b_\delta = 0, \qquad n \cdot \bar b_\delta = 0 \quad \text{on } \partial\Omega, \\ &\bar b_\delta|_{t=0} = 0, \end{aligned} \tag{6.17}$$

with $\bar v = v_1 - v_2$ and $\bar\eta_\delta = \eta_{1,\delta} - \eta_{2,\delta}$, we apply Proposition 5.3 to obtain

$$[\nabla \bar b_\delta]_{r,T} \leqq K([\bar\eta_\delta\,\mathrm{rot}\,b_{2,\delta}]_{r,T} + [\bar v]_{r,T}) \tag{6.18}$$

for $2 \leqq r < p_0(\alpha/\beta)$. For $\bar v$ we have

$$\begin{aligned} &\partial_t \bar v = \Delta \bar v - B\Lambda^*\,\mathrm{rot}\,\bar b - \mathrm{grad}\,p, \qquad \mathrm{div}\,\bar v = 0 \quad \text{in } \Omega, \\ &\bar v = 0 \quad \text{on } \partial\Omega, \\ &\bar v|_{t=0} = 0 \quad \text{on } \Omega. \end{aligned} \tag{6.19}$$

We multiply the first equations of (6.17) and (6.19) with $\bar{b}_\delta$ and $\bar{v}_\delta$, respectively, add two identities and integrate over Ω. Integrating by parts using boundary conditions, we obtain

$$\frac{d}{2dt}(\|\bar{b}_\delta\|_2^2+\|\bar{v}\|_2^2)+\|\nabla\bar{v}\|_2^2(t)+\int_\Omega \eta_{1,\delta}|\nabla\bar{b}_\delta|^2\,dx$$

$$=-\int_\Omega \bar{b}_\delta\bar{\eta}_\delta \text{ rot } b_{2,\delta}\,dx+\int_\Omega (\bar{b}_\delta \text{ rot } {}^*(\bar{v}\Lambda B)-\bar{v}(B\Lambda^* \text{ rot } \bar{b}))\,dx.$$

Integrating over $(0, t)$ and applying the Cauchy inequality to the right-hand side yields

$$\|\bar{b}_\delta\|_2^2(t)+\|\bar{v}\|_2^2(t)+2[\nabla\bar{v}]_{2,t}+\alpha[\nabla\bar{b}_\delta]_{2,t}$$

$$\leqq C_\alpha[\bar{\eta}_\delta \text{ rot } b_{2,\delta}]_{2,t} \tag{6.20}$$

$$+\int_0^t\int_\Omega (\bar{b}_\delta \text{ rot } {}^*(\bar{v}\Lambda B)-\bar{v}(B\Lambda^* \text{ rot } \bar{b}))\,dx\,ds, \qquad 0\leqq t\leqq T,$$

with C_α depending only on α. By (6.16) and integrating by parts we observe that the last term in (6.20) tends to zero as $\delta\to 0$. Letting $\delta\to 0$ in (6.20) yields *an energy inequality*

$$\sup_{0\leqq t\leqq T}(\|\bar{b}\|_2^2(t)+\|\bar{v}\|_2^2(t))+2[\nabla\bar{v}]_{2,T}+\alpha[\nabla\bar{b}]_{2,T}\leqq C_\alpha[\bar{\eta} \text{ rot } b_2]_{2,T}. \tag{6.21}$$

Similarly letting $\delta\to 0$ in (6.18) yields

$$[\nabla\bar{b}]_{r,T}\leqq K([\bar{\eta} \text{ rot } b_2]_{r,T}+[\bar{v}]_{r,T}). \tag{6.22}$$

If $r\leqq 10/3$, (6.21) together with (6.12a, b) yields

$$[\bar{v}]_{r,T}\leqq C[\bar{\eta} \text{ rot } b_2]_{2,T},$$

with C depending only on α, Ω, and T. Since we have assumed either (a) or (b) of Proposition 6.2, we may conclude from (6.22) that

$$[\nabla\bar{b}]_{r,T}\leqq C'[\bar{\eta} \text{ rot } b_2]_{r,T} \tag{6.23}$$

for $r>N$ sufficiently close to N with $C'=C'(\alpha,\beta,T,\Omega,r)$.

An a priori estimate for the Stokes system (6.19) (see [17]) and (6.23) yields

$$[\nabla^2\bar{v}]_{r,T}\leqq C_r[\nabla\bar{b}]_{r,T}\leqq C_rC'[\bar{\eta} \text{ rot } b_2]_{r,T}\leqq C_rC'[\bar{\eta}]_{p,T}[\text{rot } b_2]_{r+\varepsilon,T}, \tag{6.24}$$

with $1/p+1/(r+\varepsilon)=1/r$, $1<p<\infty$, $\varepsilon>0$. Since $r>N$ sufficiently close to N, (6.8) and (6.12a, b) imply that $[\text{rot } b_2]_{r+\varepsilon,T}\leqq Z''$ where Z'' has the same property as Z in (6.11). By Proposition 4.4 we conclude that

$$[\bar{\zeta}]_{1,T}\leqq TL\int_0^T\|\bar{v}\|_\infty(s)\,ds\leqq C^*TL\int_0^T\|\nabla^2\bar{v}\|_r(s)\,ds\leqq C^*T^{2-1/r}L[\nabla^2\bar{v}]_{r,T}$$

by the Sobolev inequality with some constant C^*. Applying (6.24) yields

$$[\bar{\zeta}]_{1,T}\leqq C_rC'C^*LT^{2-1/r}Z''[\bar{\eta}]_{p,T},$$

which is the same as (6.15) by setting $c=C_rC'C^*LT^{2-1/r}Z''$. Since

$$[\bar{\zeta}]_{\rho,T}\leqq[\bar{\zeta}]_{1,T}^{1/\rho}[\bar{\zeta}]_{\infty,T}^{1-1/\rho}\leqq(2\beta)^{1-1/\rho}[\bar{\zeta}]_{1,T}^{1/\rho},$$

it follows from (6.15) that

$$[\bar{\zeta}]_{\rho,T}\leqq(2\beta)^{1-1/\rho}c^{1/\rho}[\bar{\eta}]_{p,T}^{1/\rho}. \tag{6.25}$$

If $p \leqq \rho$, this implies that F is continuous in $L^\rho(Q_T)$. If $p \geqq \rho$, since

$$[\bar{\eta}]_{p,T} \leqq [\bar{\eta}]^{\theta}_{\rho,T} [\bar{\eta}]^{1-\theta}_{\infty,T} \leqq [\bar{\eta}]^{\theta}_{\rho,T} (2\beta)^{1-\theta}$$

with $1/p = \theta/\rho$, (6.25) implies the continuity of F in $L^\rho(Q_T)$. □

Proof of the Main Existence Theorem. By Propositions 6.3–6.5 we apply the Schauder fixed-point theorem to F, and observe that there is $\eta \in S(L)$ such that $F(\eta) = \eta$ provided that L is sufficiently large. Since $(b, v) = (G_1(\eta), G_2(\eta))$, Propositions 6.1 and 6.2 together with (6.12a, b) imply that (b,v) has the desired regularity properties and the energy inequality holds. Since $E(G_2(\eta)) = \eta$, we see (b, v, η) solves a weak form of (2.9)–(2.12) with initial conditions, which completes the proof. □

REFERENCES

[1] B. BATEMAN, MHD *Instabilities*, MIT Press, Cambridge, MA, 1978.

[2] C. BARDOS, A. Y. LEROUX, AND J. C. NEDELEC, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.

[3] L. CAFFARELLI, R. KOHN, AND L. NIRENBERG, *Partial regularity of suitable weak solutions of the Navier–Stokes equations*, Comm. Pure Appl. Math., 35 (1982), pp. 771–831.

[4] S. CAMPANATO, *L^p regularity for weak solutions of parabolic systems*, Ann. Scuola Norm. Sup. Pisa, 7 (1980), pp. 65–85.

[5] H. FEDERER, *Geometric Measure Theory*, Springer-Verlag, Berlin, New York, 1969.

[6] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[7] D. FUJIWARA AND H. MORIMOTO, *An L_r-theorem of the Helmholtz decomposition of vector fields*, J. Fac. Sci. Univ. Tokyo, Sect. IA. Math. 24 (1977), pp. 685–700.

[8] Y. GIGA, *Analyticity of the semigroup generated by the Stokes operator in L_r spaces*, Math. Z., 178 (1981), pp. 297–329.

[9] Y. GIGA AND T. MIYAKAWA, *Solutions in L_r of the Navier–Stokes initial value problem*, Arch. Rational Mech. Anal., 89 (1985), pp. 267–281.

[10] E. GIUSTI, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Boston, 1984.

[11] S. N. KRUZKOV, *First-order quasilinear equations in several independent variables*, Math. USSR-Sb., 10 (1970), pp. 217–243.

[12] O. A. LADYZHENSKAYA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and quasilinear equations of parabolic type*, Trans. Amer. Math. Soc., 23 (1968).

[13] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications* I, Springer-Verlag, Berlin, New York, 1972.

[14] ———, *Problemi ai limiti non omogenei* (V), Ann. Scuola Norm. Sup. Pisa, 16 (1962), pp. 1–44. (In Italian.)

[15] T. MIYAKAWA, *A kinetic approximation of entropy solutions of first-order quasilinear equations*, in Recent Topics in Nonlinear PDE, M. Mimura and T. Nishida, eds., North-Holland, Amsterdam, New York, 1983, pp. 93–105.

[16] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics*, Vol. II, Academic Press, New York, 1975.

[17] V. A. SOLONNIKOV, *Estimates for solutions of nonstationary Navier–Stokes equations*, J. Soviet Math., 8 (1977), pp. 467–523.

[18] R. TEMAM, *Navier–Stokes Equations*, North-Holland, Amsterdam, New York, 1977.

[19] A. I. VOL'PERT, *The space* BV *and quasilinear equations*, Math. USSR-Sb., 2 (1967), pp. 225–267.

[20] Z. YOSHIDA AND Y. GIGA, *On the Ohm–Navier–Stokes system in magnetohydrodynamics*, J. Math. Phys., 24 (1983), pp. 2860–2864.

SIAM J. MATH ANAL.
Vol. 21, No. 5, pp. 1139–1171, September 1990
©1990 Society for Industrial and Applied Mathematics
003

STABILITY OF PLANAR WAVE SOLUTIONS TO A COMBUSTION MODEL*

DAVID TERMAN†

Abstract. A system of reaction diffusion equations which arise as a model for a one-step combustion process is considered. The primary concern is with the stability of planar wave solutions to this model. This problem has been studied extensively from the point of view of matched asymptotics in the limit of infinite activation energy. The asymptotic analysis has demonstrated that for a large range of parameters, the planar wave solution is unstable. As a particular parameter is varied, the planar wave solution may undergo either a Hopf or steady state bifurcation. This paper gives a rigorous mathematical justification of some of the asymptotic results.

Key words. reaction-diffusion equations, planar wave equations, high activation energy

AMS(MOS) subject classifications. 35B35, 35K55, 35K57

1. Introduction. We consider the thermodiffusive model for a premixed flame arising from the one-step chemical reaction $A \to B$. This model gives rise to the system of reaction-diffusion equations:

$$\begin{aligned} T_t &= \Delta T + QYf(T) \\ Y_t &= \Lambda^{-1}\Delta Y - Yf(T) \end{aligned} \tag{1.1}$$

where $T(x,y,t)$ is the temperature, $Y(x,y,t)$ is the mass fraction of the reactant A, $(x,y,t) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$, Δ is the usual Laplace operator with respect to the spatial variables x and y, and Q, the heat release of the reaction, and Λ, the Lewis number, are both assumed to be positive. For $f(T)$ we assume that there exists $T_* > 0$ such that

$$\begin{aligned} &\text{(a)} \quad f(T) = 0 \quad \text{if } 0 \le T < T_* \\ &\text{(b)} \quad f(T) = B_0 e^{-E/RT} \quad \text{for } T \ge T_* . \end{aligned} \tag{1.2}$$

Here B_0, R, and E are all assumed to be positive constants. The activation energy, E, will be taken to be very large. The constant T_* is often referred to as an ignition temperature; for $T < T_*$ no reaction can take place. For $T > T_*$, $f(T)$ is the usual Arrhenius kinetic term.

By a planar wave solution of (1.1) we mean a nonconstant, bounded solution of the form

$$(T(x,y,t), Y(x,y,t)) = (T_0(z), Y_0(z)) \tag{1.3}$$

where $z = (x,y) \cdot \nu + \theta t$ for some unit vector $\nu \in \mathbb{R}^2$, and $\theta > 0$ is the wave speed. We assume, without loss of generality, that $\nu = (1,0)$, so that the wave is moving in the negative x direction. The planar wave solution is assumed to satisfy boundary conditions of the form

$$\lim_{z \to -\infty} (T_0(z), Y_0(z)) = (T_-, Y_-) \quad \text{and} \quad \lim_{z \to +\infty} (T_0(z), Y_0(z)) = (T_+, 0). \tag{1.4}$$

*Received by the editors January 30, 1989; accepted for publication October 25, 1989. The work of the author was supported in part by National Science Foundation grant DMS-8702693.

†Department of Mathematics, Ohio State University, Columbus, Ohio 43210.

The unburned state (T_-, Y_-) is prescribed with $0 \leq T_- < T_*$ and $Y_- > 0$. This actually determines the burned state as $T_+ = T_- + QY_-$; see [1].

Our primary concern is with the stability of the planar wave solutions. This problem has been studied extensively from the point of view of matched asymptotics in the limit of infinite activation energy. See Sivashinsky [9], Joulin and Linan [7], Fife [4], Margolis and Matkowsky [8], and the references cited there. The asymptotic analysis has demonstrated that, for a large range of parameter values, the planar wave solutions are unstable. As a particular parameter, the Lewis number, for example, is varied, the planar wave may undergo either a Hopf or steady state bifurcation. In this paper we give a rigorous, mathematical justification of some of the asymptotic results.

Before stating our results it will be convenient to first rescale the equations (1.1) in a standard way. See [1], for example. Let ϵ be the reciprocal of the Zeldovich number; that is

$$\frac{1}{\epsilon} = \frac{E}{RT_+}\frac{T_+ - T_-}{T_+}.$$

Then let

$$\hat{T} = \frac{T - T_-}{T_+ - T_-}, \qquad \hat{Y} = \frac{Y}{Y_-},$$
$$\alpha = \frac{T_+ - T_-}{T_+}, \qquad \beta = \epsilon(B_0 e^{-E/RT_+})^{1/2},$$
$$\hat{t} = \beta^2 t, \quad \hat{x} = \beta x, \quad \hat{y} = \beta y.$$

Then (1.1) becomes, after dropping the hats,

$$\begin{aligned} T_t &= \Delta T + \frac{1}{\epsilon^2} Y \phi\left(\frac{T-1}{\epsilon}, \epsilon\right) \\ Y_t &= \Lambda^{-1}\Delta Y - \frac{1}{\epsilon^2} Y \phi\left(\frac{T-1}{\epsilon}, \epsilon\right) \end{aligned} \tag{1.5}$$

where, if $s = (T-1)/\epsilon$, then $\phi(s,\epsilon) = 0$ for $0 \leq T < T_*$, and $\phi(s,\epsilon) = \exp(s/1 + \epsilon\alpha s)$ for $T \geq T*$.

A planar wave solution is then a solution of (1.5) of the form

$$(T(x,y,t), Y(x,y,t)) = (T_\epsilon(z), Y_\epsilon(z)), \qquad z = x + \theta_\epsilon t, \tag{1.6}$$

where

$$\lim_{z\to-\infty}(T_\epsilon(z), Y_\epsilon(z)) = (0,1) \quad \text{and} \quad \lim_{z\to\infty}(T_\epsilon(z), Y_\epsilon(z)) = (1,0). \tag{1.7}$$

Note that a planar wave solution satisfies the system of ordinary differential equations

$$\begin{aligned} T_\epsilon'' - \theta_\epsilon T_\epsilon' + \frac{1}{\epsilon^2} Y_\epsilon \phi\left(\frac{T_\epsilon - 1}{\epsilon}, \epsilon\right) &= 0, \\ \Lambda^{-1} Y_\epsilon'' - \theta_\epsilon Y_\epsilon' - \frac{1}{\epsilon^2} Y_\epsilon \phi\left(\frac{T_\epsilon - 1}{\epsilon}, \epsilon\right) &= 0, \end{aligned} \tag{1.8}$$

together with the boundary conditions (1.7).

Berestycki, Nicolaenko, and Scheurer [1] proved that for each $\epsilon > 0$, there exists a solution of (1.7), (1.8) for some speed θ_ϵ. They also analyzed the asymptotic behavior

of the solution as $\epsilon \to 0$. They demonstrated that $\lim_{\epsilon \to 0} \theta_\epsilon = \sqrt{2\Lambda}$, and the solution converges strongly to a limiting free boundary problem. In their paper, Berestycki, Nicolaenko, and Scheurer assumed, for simplicity, that $\phi(s, \epsilon)$ does not depend on ϵ with the most important case being $\phi(s, \epsilon) = \phi(s) = e^s$ for $T > T_*$. In this paper we also make this assumption; that is, we assume that $\phi(s, \epsilon) = \phi(s)$ where

$$\begin{aligned} &\text{(a)} \quad \phi_\epsilon\left(\frac{T-1}{\epsilon}\right) = 0 \quad \text{if } T < T_* \\ &\text{(b)} \quad \phi_\epsilon\left(\frac{T-1}{\epsilon}\right) = \exp\left(\frac{T-1}{\epsilon}\right) \quad \text{if } T \geq T_*. \end{aligned} \tag{1.9}$$

The purpose of this assumption is to simplify the (already tedious) calculations which follow. The results of this paper still hold for the original equations.

We now consider the stability of the planar wave solution. We linearize the equations (1.5) about the planar wave solution and then compute the eigenvalues of the resulting equations. The first step is to change to a moving coordinate system; that is, we let $z = x + \theta_\epsilon t$. We then linearize the resulting equations about the planar wave solution to obtain the system

$$\begin{aligned} \hat{u}_t &= \Delta\hat{u} - \theta_\epsilon \hat{u}_z + \frac{1}{\epsilon^3} Y_\epsilon \phi\left(\frac{T_\epsilon - 1}{\epsilon}\right)\hat{u} + \frac{1}{\epsilon^2}\phi\left(\frac{T_\epsilon - 1}{\epsilon}\right)\hat{v}, \\ \hat{v}_t &= \Lambda^{-1}\Delta\hat{v} - \theta_\epsilon \hat{v}_z - \frac{1}{\epsilon^3} Y_\epsilon \phi\left(\frac{T_\epsilon - 1}{\epsilon}\right)\hat{u} - \frac{1}{\epsilon^2}\phi\left(\frac{T_\epsilon - 1}{\epsilon}\right)\hat{v}. \end{aligned} \tag{1.10}$$

Here, $\hat{u}$ and $\hat{v}$ represent, respectively, the variations in T and Y. They are functions of z, y, and t, and Δ is now the Laplace operator with respect to the variables z and y. We look for a solution of (1.10) of the form

$$\hat{u}(z, y, t) = e^{iky + \sigma t} u(z) \quad \text{and} \quad \hat{\mathrm{v}}(\mathrm{z}, \mathrm{y}, \mathrm{t}) = \mathrm{e}^{\mathrm{iky} + \sigma \mathrm{t}} \mathrm{v}(\mathrm{z})$$

where $k \geq 0$ and σ is complex. Hence, we are looking for perturbations of (1.5) which are periodic in y with wave number k, and which grow (or decay) in time at the rate $e^{\sigma t}$. Note that $(u(z), v(z))$ satisfy the system of ordinary differential equations:

$$\begin{aligned} u'' - \theta_\epsilon u' + \frac{1}{\epsilon^3} Y_\epsilon \phi\left(\frac{T_\epsilon - 1}{\epsilon}\right) u + \frac{1}{\epsilon^2}\phi\left(\frac{T_\epsilon - 1}{\epsilon}\right) v &= (\sigma + k^2) u, \\ \Lambda^{-1} v'' - \theta_\epsilon v' - \frac{1}{\epsilon^3} Y_\epsilon \phi\left(\frac{T_\epsilon - 1}{\epsilon}\right) u - \frac{1}{\epsilon^2}\phi\left(\frac{T_\epsilon - 1}{\epsilon}\right) v &= (\sigma + \Lambda^{-1} k^2) v \end{aligned} \tag{1.11}$$

if $T_\epsilon \neq T^*$. If $T_\epsilon = T_*$, then $\phi(T_\epsilon - 1/\epsilon)$ is discontinuous. Hence, we cannot expect $u(z)$ and $v(z)$ to be differentiable whenever $T_\epsilon(z) = T_*$. Berestycki, Nicolaenko, and Scheurer [1] proved that $T_\epsilon'(z) > 0$ for all z. Hence, there exists a unique z^ϵ such that $T_\epsilon(z^\epsilon) = T_*$. We shall impose the natural regularity conditions on u and v. We assume that $u(z)$ and $v(z)$ are both continuous for all z, twice continuously differentiable if $z \neq z^\epsilon$, and

$$u'(z^\epsilon +) - u'(z^\epsilon -) = -\omega, \qquad v'(z^\epsilon +) - v'(z^\epsilon -) = \Lambda\omega \tag{1.12}$$

where

$$\omega = \frac{Y_\epsilon(z^\epsilon) u(z^\epsilon) \exp(\mathrm{T}_* - 1/\epsilon)}{\epsilon^3 \mathrm{T}'(\mathrm{z}^\epsilon)}.$$

DEFINITION. Assume that $k > 0$. The planar wave solution of (1.5) is linearly unstable (stable) to perturbations with wave numbers k, if there exists (does not exist) a bounded, nontrivial solution of (1.11) with $\operatorname{Re}\sigma > 0$ ($\operatorname{Re}\sigma \geq 0$). The case $k = 0$ will be discussed shortly.

Before stating our main results we set

$$\begin{aligned} &\Lambda = 1 + \epsilon\ell, \qquad \Gamma = \sqrt{1 + 2(\sigma + k^2)}\,, \\ &D(\sigma, k, \ell) = 2\Gamma^2(\Gamma - 1) - \ell(\sigma - \Gamma + 1)\,. \end{aligned} \tag{1.13}$$

We assume throughout this paper that

(1.14)

(a) $\varepsilon > 0, \quad k \geq 0, \quad \Lambda = 1 + \varepsilon\ell > 0,$

(b) $\operatorname{Re}\left(\theta_\epsilon - \sqrt{\theta_\epsilon^2 + 4(\sigma + k^2)}\,\right) < 0$ and $\operatorname{Re}\left(\theta_\epsilon - \sqrt{\theta_\epsilon^2 + 4(\sigma + \Lambda k^2)}\,\right) < 0.$

Remarks concerning this last assumption will be given later.

PROPOSITION 1.1. *Assume* (1.14) *is satisfied. There exists a complex valued function* $D_\epsilon(\sigma, k, \ell)$, *which is holomorphic in* σ, *and has the property that there exists a bounded, nontrivial solution of* (1.11) *if and only if* $D_\epsilon(\sigma, k, \ell) = 0$.

In the following theorem we compute $D_\epsilon(\sigma, k, \ell)$ in the limit as $\epsilon \to 0$.

THEOREM 1.2. *Assume that* (1.14) *is satisfied. Then* $\lim_{\epsilon \to 0} D_\epsilon(\sigma, k, \ell) = D(\sigma, k, \ell)$.

Remark 1. The complex valued function $D(\sigma, k, \ell)$ was obtained by Sivashinsky [9], Joulin and Linan [7], Fife [4], and others using the methods of matched asymptotics in the limit $\epsilon = 0$. Since $D_\epsilon(\sigma k, \ell)$ is defined for ϵ positive but small (or E finite but large), Theorem 1.2 gives a rigorous justification of these previous asymptotic results.

Remark 2. Recall Rouche's theorem which states that if two holomorphic functions are close to each other, then so are their roots. Hence, if (1.14) is satisfied and ϵ is sufficiently small, then the values of σ for which there exists a bounded solution of (1.11) are close to the roots of $D(\sigma, k, \ell)$. In particular, we have the following result.

COROLLARY 1.3. *Fix* k *and* ℓ. *If there exists a root of* $D(\sigma, k, \ell)$ *with* $\operatorname{Re}\sigma > 0$, *then, for* ϵ *sufficiently small, the planar wave solution is linearly unstable to perturbations with wave number* k.

It is a simple manner to determine the roots of $D(\sigma, k, \ell)$. See [8].

PROPOSITION 1.4. *There exist two smooth curves,* $\ell = h_1(k) = -2 - 2k^2$ *and* $\ell = h_2(k)$, *such that for each* $k > 0$,

(a) $h_1(k) < 0 < h_2(k)$, $\lim_{k \to \infty} h_1(k) = -\infty$, $\lim_{k \to \infty} h_2(k) = +\infty$,

(b) *If either* $\ell < h_1(k)$ *or* $\ell > h_2(k)$, *then there exists* σ *with* $\operatorname{Re}\sigma > 0$ *such that* $D(\sigma, k, \ell) = 0$,

(c) *If* $h_1(k) < \ell < h_2(k)$, *then there does not exist* σ *with* $\operatorname{Re}\sigma \geq 0$ *such that* $D(\sigma, k, \ell) = 0$,

(d) *If* $\ell = h_1(k)$, *then* $D(0, k, \ell) = 0$,

(e) *If* $\ell = h_2(k)$, *then there exists* σ *with* $\operatorname{Re}\sigma = 0$ *and* $\operatorname{Im}\sigma \neq 0$, *such that* $D(\sigma, k, \ell) = D(\bar{\sigma}, k, \ell) = 0$.

If $k = 0$, *then* $D(0, k, \ell) = 0$. *Moreover, there exists* σ *with* $\operatorname{Re}\sigma > 0$ *such that* $D(\sigma, 0, \ell) = 0$ *if and only if* $\ell > h_2(k)$.

Theorem 1.2, together with the Implicit Function Theorem demonstrates that the results of Proposition 1.4 perturb for $\epsilon > 0$.

THEOREM 1.5. *Fix $M > 0$. There exists $\epsilon_0 > 0$ such that if $0 < \epsilon < \epsilon_0$, then there exist smooth curves $\ell = h_1^\epsilon(k)$ and $\ell = h_2^\epsilon(k)$, defined for $0 \le k \le M$, such that*

(a) $\lim_{\epsilon \to 0} h_i^\epsilon(k) = h_i(k), \quad i = 1, 2,$

(b) *If $k > 0$ and either $-M < \ell < h_1^\epsilon(k)$ or $h_2^\epsilon(k) < \ell < M$, then there exists σ with $\operatorname{Re}\sigma > 0$ such that $D_\epsilon(\sigma, k, \ell) = 0$,*

(c) *If $k > 0, h_1^\epsilon(k) < \ell < h_2^\epsilon(k), |\sigma| < M$, and $\operatorname{Re}\sigma > 0$, then $D_\epsilon(\sigma, k, \ell) \neq 0$,*

(d) *If $k > 0$ and $\ell = h_1^\epsilon(k)$, then $D_\epsilon(0, k, \ell) = 0$,*

(e) *If $k \ge 0$ and $\ell = h_2^\epsilon(k)$, then there exists σ with $\operatorname{Re}\sigma = 0$ and $\operatorname{Im}\sigma \neq 0$ such that $D_\epsilon(\sigma, k, \ell) = D_\epsilon(\overline{\sigma}, k, \ell) = 0$,*

(f) *If $k = 0$, then $D_\epsilon(0, k, \ell) = 0$. If $\ell > h_2^\epsilon(0)$, then there exists σ with $\operatorname{Re}\sigma > 0$ such that $D_\epsilon(\sigma, 0, \ell) = 0$. If $\ell < h_2^\epsilon(0), |\sigma| < M, \operatorname{Re}\sigma \ge 0$, and $\sigma \neq 0$, then $D_\epsilon(\sigma, 0, \ell) \neq 0$.*

Remark. In parts (c) and (f) we assume that $|\sigma| < M$. Because of this assumption we are unable to conclude that the planar wave solution is linearly stable for $h_1^\epsilon(k) < \ell < h_2^\epsilon(k)$ if $k > 0$, and for $\ell < h_2^\epsilon(k)$ if $k = 0$.

It is straightforward to show that if the planar wave solution is linearly unstable, then it is really unstable as a solution of (1.5). It is not clear, however, if linear stability implies nonlinear stability. This will be true if we restrict ourselves to perturbations which are constant in the y-direction. In order to state this result precisely, we may, without loss of generality, drop the dependence on the y variable entirely in (1.5) and (1.10).

For ϵ sufficiently small, we consider the linear operator,

$$L_\epsilon \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u'' - \theta_\epsilon u' + \dfrac{1}{\epsilon^3} Y_\epsilon \varphi\left(\dfrac{T_\epsilon - 1}{\epsilon}\right) u + \dfrac{1}{\epsilon^2}\varphi\left(\dfrac{T_\epsilon - 1}{\epsilon}\right) v \\ \Lambda^{-1} v'' - \theta_\epsilon v' - \dfrac{1}{\epsilon^3} Y_\epsilon \varphi\left(\dfrac{T_\epsilon - 1}{\epsilon}\right) u - \dfrac{1}{\epsilon^2}\varphi\left(\dfrac{T_\epsilon - 1}{\epsilon}\right) v \end{bmatrix}$$

defined on the weighted Banach space

$$X_\rho = \{(u_1(z), u_2(z)) \mid u_i \text{ is uniformly continuous and } \sup|e^{\rho|z|} u_i(z)| < \infty \quad \text{for i = 1 and 2}\}.$$

Here ρ is fixed with $0 < \rho < \theta_\epsilon$. The results of Henry [6, p. 140] imply that L_ϵ defines a sectorial operator with essential spectrum in the left half complex plane bounded away from the imaginary axis. From the definitions given in §3, we find that in the weighted space X_ρ, the holomorphic function $D_\epsilon(\sigma, 0, \ell)$ is well defined in an open region which contains the right half complex plane. From Theorem 1.2 and the explicit formula for $D(\sigma, k, \ell)$ given in (1.13), we easily conclude that $\sigma = 0$ is a simple eigenvalue for ϵ sufficiently small. The results of Henry [6] now imply that the planar wave solution is stable to the one-dimensional perturbations if there does not exist an eigenvalue σ of the linear operator L_ϵ with $\operatorname{Re}\sigma \ge 0$, $\sigma \neq 0$. Therefore, if we could obtain an a priori bound, independent of ϵ, on the eigenvalues of L_ϵ, then we would be able to conclude that the planar wave solution is stable to one-dimensional perturbations for $\ell < h_2^\epsilon(0)$.

We also point out that two complex eigenvalues cross the imaginary axis as the parameter ℓ increases past $h_\epsilon^2(0)$. Theorem 1.2, together with the explicit formula for $D(\sigma, k, \ell)$ given in (1.3), implies that this crossing is transversal if ϵ is sufficiently small. Henry's results now imply that if we are able to obtain an a priori bound on the eigenvalues of L_ϵ, then a Hopf bifurcation takes place at $\ell = h_2^\epsilon(0)$.

The organization of the paper is as follows. In §2 we prove the estimates needed on the planar wave solutions $(T_\epsilon(z), Y_\epsilon(z))$. Certainly we need to understand these functions very well, because they appear explicitly in the linear equations (1.11). We prove Proposition 1.1 in §3. We begin the proof of Theorem 1.2 in §4, with the explicit computation of $D_\epsilon(\sigma, k, \ell)$ given in §5. The technical details of the proof of Theorem 1.2 are carried out in §6.

We assume throughout this paper that (1.14) is satisfied and $M > 0$ is chosen so that

$$\begin{aligned}
&\text{(a)} \quad |\ell| < \frac{1}{2}M, \\
&\text{(b)} \quad 0 \le k \le M, \\
&\text{(c)} \quad |\sigma| < M \\
&\text{(d)} \quad \operatorname{Re}\left(\theta_\epsilon - \sqrt{\theta_\epsilon^2 + 4(\sigma + \mathrm{K}^2)}\right) < -\frac{1}{\mathrm{M}}, \\
&\text{(e)} \quad \operatorname{Re}\left(\theta_\epsilon - \sqrt{\theta_\epsilon^2 + 4(\sigma + \Lambda\mathrm{k}^2)}\right) < -\frac{1}{\mathrm{M}}.
\end{aligned} \tag{1.15}$$

2. Estimates for the wave.

A. Introduction. In this section we present the estimates needed on the planar wave solution for the remainder of the paper. Many of these estimates are extensions of the results of Berestycki, Nicolaenko, and Scheuer [1] who proved the existence of the planar wave solution. They also proved that there exists a positive constant K such that if ϵ is sufficiently small, then for each z,

$$\begin{aligned}
&\text{(a)} \quad 0 < T_\epsilon(z) < 1, \qquad 0 < Y_\epsilon(z) < 1, \\
&\text{(b)} \quad 0 < T_\epsilon'(z) < K, \qquad -K < Y_\epsilon'(z) < 0, \\
&\text{(c)} \quad |T_\epsilon(z) + \Lambda^{-1} Y_\epsilon(z) - 1| \le |1 - \Lambda^{-1}|(1 - T_\epsilon(z)).
\end{aligned} \tag{2A.1}$$

Note that every translation of a planar wave solution is also a planar wave solution. Throughout this paper we fix the translation so that

$$T_\epsilon(0) = 1 + 6\epsilon \ln \epsilon. \tag{2A.2}$$

We choose z^ϵ and z_ϵ so that

$$T_\epsilon(z^\epsilon) = T_* \quad \text{and} \quad \mathrm{T}_\epsilon(\mathrm{z}_\epsilon) = 1 - \epsilon. \tag{2A.3}$$

We derive estimates in each of the intervals $z \le z^\epsilon$, $z^\epsilon < z \le 0$, $0 < z \le z_\epsilon$, and $z_\epsilon < z$ separately.

B. $\mathbf{z_\epsilon < z}$. It will be convenient to introduce the variables

$$p_\epsilon = \frac{T_\epsilon - 1}{\epsilon}, \quad q_\epsilon = \frac{Y_\epsilon}{\epsilon}, \quad \xi = \frac{z}{\epsilon}. \tag{2B.1}$$

Let $\xi_\epsilon = \left(\frac{1}{\epsilon}\right) z_\epsilon$. In these new variables (1.8) becomes

$$\begin{aligned}
&p_\epsilon'' - \epsilon\theta_\epsilon p_\epsilon' + q_\epsilon \phi(p_\epsilon) = 0, \\
&\Lambda^{-1} q_\epsilon'' - \epsilon\theta_\epsilon q_\epsilon' - q_\epsilon \phi(p_\epsilon) = 0
\end{aligned} \tag{2B.2}$$

where differentiation is with respect to ξ.

We now state the main results of this section. The proofs of these results are then split into a number of lemmas which follow.

PROPOSITION 2B.1. *There exist positive constants α and β, which do not depend on ϵ, such that if ϵ is sufficiently small, then for $\xi \geq \xi_\epsilon$,*

$$\text{(2B.3)} \qquad \begin{aligned} &\text{(a)} \quad e^{-\beta(\xi-\xi_\epsilon)} < |p_\epsilon(\xi)| < e^{-\alpha(\xi-\xi_\epsilon)}, \\ &\text{(b)} \quad \tfrac{1}{2}e^{-\beta(\xi-\xi_\epsilon)} < |q_\epsilon(\xi)| < 2e^{-\alpha(\xi-\xi_\epsilon)}. \end{aligned}$$

For the next proposition we introduce the local Shvab–Zeldvich variable

$$\text{(2B.4)} \qquad \Phi_\epsilon(\xi) = p_\epsilon(\xi) + \Lambda^{-1}q_\epsilon(\xi).$$

PROPOSITION 2B.2. *There exists a positive constant K_1, which does not depend on ϵ, such that if ϵ is sufficiently small, then for $\xi \geq \xi_\epsilon$, $\max\{|\Phi_\epsilon(\xi)|, |\Phi'_\epsilon(\xi)|\} \leq \epsilon^2 K_1 e^{-\alpha(\xi-\xi_\epsilon)}$.*

Here, α is as in the preceding proposition. Using the definitions, we conclude from Proposition 2B.2 that

$$\text{(2B.5)} \qquad |T_\xi(\xi) + \Lambda^{-1}Y_\xi(\xi)| \leq \epsilon^3 K_1 e^{-\alpha(\xi-\xi_\epsilon)}.$$

The proofs of the preceding propositions are now split into a number of lemmas.

LEMMA 2B.3. *If ϵ is sufficiently small, then $-\frac{1}{2}p_\epsilon(\xi) < q_\epsilon(\xi) < -2p_\epsilon(\xi)$ for all ξ.*

Proof. If we divide (2A.1c) by ϵ, and use (1.15), then the result follows easily.

LEMMA 2B.4. *There exists η_0 such that if ϵ is sufficiently small, and $\xi > \eta_0$, then $-\frac{1}{2}p_\epsilon(\xi) < p'_\epsilon(\xi) < -2p_\epsilon(\xi)$.*

Proof. We begin by writing (2B.2) as a first-order system:

$$\text{(2B.6)} \qquad \begin{aligned} p'_\epsilon &= y_\epsilon, \\ y'_\epsilon &= \epsilon\theta_\epsilon y_\epsilon - q_\epsilon\phi(p_\epsilon) \\ q'_\epsilon &= w_\epsilon, \\ w'_\epsilon &= \epsilon\theta_\epsilon\Lambda w_\epsilon + q_\epsilon\Lambda\phi(p_\epsilon). \end{aligned}$$

To determine the asymptotic behavior of $p_\epsilon(\xi)$ as $\xi \to \infty$, we linearize (2B.6) at the origin, O, to obtain the system

$$W'_\epsilon = A_\epsilon W_\epsilon$$

where

$$W_\epsilon = \begin{pmatrix} p_\epsilon \\ y_\epsilon \\ q_\epsilon \\ w_\epsilon \end{pmatrix} \quad \text{and} \quad \mathrm{A}_\epsilon = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & \epsilon\theta & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \Lambda & \epsilon\theta_\epsilon\Lambda \end{bmatrix}.$$

The matrix A_ϵ has two positive and one negative eigenvalues. There is also one zero eigenvalue due to the fact that the line $q_\epsilon = 0$ consists of rest points of (2B.2). The negative eigenvalue of A_ϵ, and a corresponding eigenvector are

$$m_\epsilon = \frac{\epsilon\theta - \sqrt{4\Lambda + \epsilon^2\theta_\epsilon^2}}{2\Lambda} \quad \text{and} \quad \mathrm{e}_\epsilon = \begin{pmatrix} 1 \\ m_\epsilon \\ m_\epsilon(\epsilon\theta - m_\epsilon) \\ m_\epsilon^2(\epsilon\theta - m_\epsilon) \end{pmatrix}.$$

As $\epsilon \to 0$, $W_\epsilon(\xi)$ approaches O tangent to e_ϵ. Because $\lim_{\epsilon\to 0} m_\epsilon = -1$, we conclude that if ϵ is sufficiently small and z is sufficiently large, say $\xi \geq \eta_0$, then $-\frac{1}{2}p_\epsilon(\xi) < y_\epsilon(\xi) < -2p_\epsilon(\xi)$. Because $y(\xi) = p'_\epsilon(\xi)$, the result follows.

LEMMA 2B.5. *There exist constants α and β, which do not depend on ϵ, such that if ϵ is sufficiently small and $\xi \geq \xi_\epsilon$, then $-\alpha p_\epsilon(\xi) < p'_\epsilon(\xi) < -\beta p_\epsilon(\xi)$.*

Proof. As before we let $y_\epsilon(\xi) = p'_\epsilon(\xi)$. Fix $\epsilon > 0$, and α and β such that $0 < \alpha < \frac{1}{2}$ and $\beta > 2$. Let

$$\mathcal{S} = \{(p_\epsilon, y_\epsilon) : \ -1 \leq p_\epsilon < 0\,, \quad -\alpha p_\epsilon < y_\epsilon < -\beta p_\epsilon\}\,.$$

Note that $p_\epsilon(\xi_\epsilon) = -1$ and $-1 < p_\epsilon(\xi) < 0$ for $\xi > \xi_\epsilon$. We show that α and β can be chosen so that $(p_\epsilon(\xi), y_\epsilon(\xi)) \in \mathcal{S}$ for all $\xi > \xi_\epsilon$. From the preceding lemma we know that this is true for $\xi > \eta_0$. Suppose that $(p_\epsilon(\xi), y_\epsilon(\xi)) \notin \mathcal{S}$ for some $\xi \in [\xi_\epsilon, \eta_0)$. Let $\eta_1 = \sup\{\xi : \ (\mathrm{p}_\epsilon(\xi), \mathrm{y}_\epsilon(\xi)) \notin \mathcal{S}\}$. Then $(p_\epsilon(\xi), y_\epsilon(\xi))$ must be entering $\mathcal{S}$ at $\xi = \eta_1$. We shall prove that this is impossible unless $\eta_1 = \xi_\epsilon$ if α is sufficiently small and β is sufficiently large.

Suppose that $(p_\epsilon(\eta_1), y_\epsilon(\eta_1)) \in \ell_\epsilon \equiv \{(p_\epsilon, y_\epsilon) : y_\epsilon = -\alpha p_\epsilon\}$. Let $n = (\alpha, 1)$ be a vector normal to ℓ_ϵ which points into $\mathcal{S}$. Then, because $(p_\epsilon(\xi), y_\epsilon(\xi))$ is entering $\mathcal{S}$ at $\xi = \eta_1$, we must have that $n \cdot (p'_\epsilon(\eta_1), y'_\epsilon(\eta_1)) \geq 0$. However, using (2B.6) and Lemma 2B.4 we find that

$$\begin{aligned} n \cdot (p'_\epsilon(\eta_1), y'_\epsilon(\eta_1)) &= \alpha y_\epsilon(\eta_1) + \epsilon\theta_\epsilon y_\epsilon(\eta_1) - q_\epsilon(\eta_1)\phi(p_\epsilon) \\ &\leq (\alpha + \epsilon\theta_\epsilon) y_\epsilon(\eta_1) + \tfrac{1}{2} p_\epsilon(\eta_1)\phi(p_\epsilon) \\ &= -\alpha(\alpha + \epsilon\theta_\epsilon) p_\epsilon(\eta_1) + \tfrac{1}{2} p_\epsilon(\eta_1)\exp(\mathrm{p}_\epsilon(\eta_1)) \\ &< -\alpha(\alpha + \epsilon\theta_\epsilon) p_\epsilon(\eta_1) + \tfrac{1}{2} e^{-1} p_\epsilon(\eta_1) \\ &< 0 \end{aligned}$$

if α is sufficiently small, independent of ϵ. This gives the desired contradiction. A similar computation shows that if β is sufficiently large, then it is impossible for $(p_\epsilon(\eta_1), y_\epsilon(\eta_1)) \in \ell_\beta \equiv \{(p_\epsilon, y_\epsilon) : y_\epsilon = -\beta p_\epsilon\}$. This completes the proof of the lemma.

Proposition 2B.1 now follows easily from Lemmas 2B.3 and 2B.5.

Proof of Proposition 2B.2. For convenience, we shall drop the subscript ϵ. We add the equations in (2B.2) to find that

$$\Phi''(\xi) - \epsilon\theta\Phi'(\xi) = \epsilon\theta(1 - \Lambda^{-1}) q'(\xi)\,.$$

We integrate this equation from ξ to ∞ to obtain

$$\text{(2B.7)} \qquad \begin{aligned} &\text{(a)} \quad \Phi'(\xi) - \epsilon\theta\Phi_\epsilon(\xi) = \epsilon\theta(1 - \Lambda^{-1}) q(\xi)\,, \\ &\text{(b)} \quad (e^{-\epsilon\theta\xi}\Phi(\xi))' = \epsilon\theta(1 - \Lambda^{-1}) e^{-\epsilon\theta\xi} q(\xi)\,. \end{aligned}$$

We integrate this last equation from ξ to ∞ to obtain

$$\Phi_\epsilon(\xi) = \epsilon\theta(1 - \Lambda^{-1}) e^{\epsilon\theta\xi} \int_\xi^\infty e^{-\epsilon\theta\eta} q_\epsilon(\eta) d\eta\,.$$

It now follows from (1.15) and Proposition 2B.1 that

$$\begin{aligned} |\Phi_\epsilon(\xi)| &\leq 2\epsilon^2\theta M e^{\epsilon\theta\xi} \int_\xi^\infty e^{-\epsilon\theta\eta} e^{-\alpha(\eta - \xi_\epsilon)} d\eta \\ &= \frac{2\theta M}{\alpha + \epsilon\theta} \epsilon^2 e^{-\alpha(\xi - \xi_\epsilon)}\,. \end{aligned}$$

This, together with (2B.7a), implies that

$$|\Phi_\epsilon'(\xi)| \le \frac{2\theta^2 M}{\alpha+\epsilon\theta}\epsilon^3 e^{-\alpha(\xi-\xi_\epsilon)} + 2\epsilon^2\theta M e^{-\alpha(\xi-\xi_\epsilon)}$$

and the result follows.

C. $\mathbf{z \le z^\epsilon}$. If $z \le z^\epsilon$, then $(T_\epsilon(z), Y_\epsilon(z))$ satisfy

$$T_\epsilon'' - \theta_\epsilon T_\epsilon' = 0\,, \qquad \Lambda^{-1}Y_\epsilon'' - \theta_\epsilon Y_\epsilon' = 0\,,$$
$$T_\epsilon(z^\epsilon) = T_*\,, \qquad \lim_{z\to-\infty}(T_\epsilon(z), Y_\epsilon(z)) = (0,0)\,.$$

Hence, $T(z) = T_* e^{\theta_\epsilon(z-z^\epsilon)}$.

D. $\mathbf{z^\epsilon < z \le 0}$. Once again we work with the variables p_ϵ, q_ϵ, and ξ. Let $\xi^\epsilon = z^\epsilon/\epsilon$. For convenience, we shall drop the subscript ϵ throughout this section. Note that $p(0) = T(0) - 1/\epsilon = 6\ln\epsilon$. Since $p'(\xi) > 0$, it follows from Lemma 2B.3 that if $\xi \le 0$, then

$$|q\phi(p)| \le 2|pe^p| \le -12(\ln\epsilon)\epsilon^6 \le \epsilon^5 \tag{2D.1}$$

if ϵ is sufficiently small. Let $\beta(\xi) = -\frac{1}{\epsilon} + (T_*/\epsilon)e^{\epsilon\theta(\xi-\xi^\epsilon)}$ be the solution of the equations,

$$\beta'' - \epsilon\theta\beta' = 0\,, \qquad (\beta(\xi^\epsilon), \beta_\xi(\xi^\epsilon)) = (p(\xi^\epsilon), p_\xi(\xi^\epsilon))\,.$$

From (2D.1) we expect that $|p(\xi) - \beta(\xi)|$ is small for $\xi^\epsilon < \xi < 0$. This is made precise in the following proposition.

PROPOSITION 2D.1. *If ϵ is sufficiently small and $k = (\ln 2/T_*)$, then $0 < -\xi^\epsilon < k/\epsilon$. Moreover, $|p(\xi) - \beta(\xi)| < \epsilon^2$ for $\xi^\epsilon \le \xi \le 0$.*

Proof. Let $\alpha(\xi) = p(\xi) - \beta(\xi)$. Then $\alpha(\xi)$ satisfies

$$\alpha'' - \epsilon\theta\alpha' = -q\phi(p)\,, \qquad \alpha(\xi^\epsilon) = -\alpha'(\xi^\alpha) = 0\,.$$

It follows that

$$\alpha'(\xi) = -e^{\epsilon\theta(\xi-\xi^\epsilon)}\int_{\xi^\epsilon}^{\xi} e^{-\epsilon\theta(\eta-\xi^\epsilon)}q\phi(p)d\eta\,.$$

Using (2D.1) we conclude that if $\xi^\epsilon \le \xi \le 0$, then

$$|\alpha'(\xi)| \le \epsilon^5 e^{\epsilon\theta(\xi-\xi^\epsilon)}(\xi - \xi^\epsilon).$$

Suppose that $\xi - \xi^\epsilon \le \frac{k}{\epsilon} < -\xi^\epsilon$. Then, $|\alpha'(\xi)| \le k\epsilon^4 e^{\theta k}$, and

$$|\alpha(\xi)| \le \int_{\xi^\epsilon}^{\xi}|\alpha'(\eta)|d\eta \le k\epsilon^4 e^{\theta k}(\xi - \xi^\epsilon) \le k^2\epsilon^3 e^{\theta k}\,.$$

Recall that $\lim_{\epsilon\to 0}\theta_\epsilon = \sqrt{2\Lambda}$. Therefore, if ϵ is sufficiently small, then $\ln 2/T_* < \theta k < 2\ln 2/T_*$. Hence,

$$\begin{aligned} p\left(\xi^\epsilon + \frac{k}{\epsilon}\right) &= \beta\left(\xi^\epsilon + \frac{k}{\epsilon}\right) + \alpha\left(\xi^\epsilon + \frac{k}{\epsilon}\right) \\ &\ge -\frac{1}{\epsilon} + \frac{T_*}{\epsilon}e^{\theta k} - k^2\epsilon^3 e^{\theta k} \\ &\ge -\frac{1}{\epsilon} + \frac{2}{\epsilon} - k^2\epsilon^3\left(\frac{2}{T_*}\right)^2 \\ &> 0 \end{aligned}$$

if ϵ is sufficiently small. This, however, is a contradiction because $p(\xi) < 0$ for all ξ.

COROLLARY 2D.2. *If ϵ is sufficiently small, then*

$$|p'_\epsilon(0) - \theta_\epsilon(1 + 6\epsilon \ln \epsilon)| \leq \epsilon^3 .$$

Proof. Recall $\alpha(\xi)$ which was defined in the preceding proposition. If $\xi^\epsilon \leq \xi \leq 0$, then using the estimates obtained in the preceding proposition,

$$\begin{aligned} |p'(\xi) - \epsilon\theta_\epsilon p(\xi) - \theta_\epsilon| &\leq |p'(\xi) - \beta'(\xi| \\ &\quad + |\beta'(\xi) - \epsilon\theta_\epsilon\beta(\xi) - \theta_\epsilon| + \epsilon\theta_\epsilon|\beta(\xi) - p(\xi)| \\ &= |\alpha'(\xi)| + \epsilon\theta_\epsilon|\alpha(\xi)| \\ &< \epsilon^3 \end{aligned}$$

if ϵ is sufficiently small. Since $p(0) = 6 \ln \epsilon$, the result follows.

Note that $p_\xi(0) = T_x(0)$. Hence,

$$|T_z(0) - \theta_\epsilon(1 + 6\epsilon \ln \epsilon)| \leq \epsilon^3 . \tag{2D.2}$$

Moreover, using (1.8) and (2D.1), we conclude that

$$|T_{zz}(0) - \theta_\epsilon T_z(0)| \leq \epsilon^4 \quad \text{and} \quad |Y_{zz}(0) - \Lambda\theta_\epsilon Y_z(0)| \leq \epsilon^4. \tag{2D.3}$$

E. $0 \leq z \leq z_\epsilon$. It will be again more convenient to work with the variables p_ϵ, q_ϵ, and ξ. We shall often drop the subscript ϵ. Recall that $\xi_\epsilon = z_\epsilon/\epsilon$.

PROPOSITION 2E.1. *Let α be as in Proposition* 2B.1, *and $k_1 = \frac{7}{\alpha}$. If ϵ is sufficiently small, then $\xi_\epsilon < -k_1 \ln \epsilon$.*

Proof. Note that $p'' = \epsilon\theta p' - q\phi(p) \leq \epsilon\theta p'$. This implies that

$$p'(\xi) \geq p'(\xi_\epsilon) e^{-\epsilon\theta(\xi_\epsilon - \xi)} . \tag{2E.1}$$

However, from Lemma 2B.5, $p'(\xi_\epsilon) \geq -\alpha p(\xi_\epsilon) = \alpha$. Therefore,

$$\begin{aligned} p(0) &= p(\xi_\epsilon) - \int_0^{\xi_\epsilon} p'(\eta) d\eta \\ &\leq -1 - \alpha \int_0^{\xi_\epsilon} e^{-\epsilon\theta(\xi_\epsilon - \eta)} d\eta \\ &= -1 - \frac{\alpha}{\epsilon\theta}\left[1 - e^{-\epsilon\theta\xi_\epsilon}\right] . \end{aligned}$$

If $\xi_\epsilon \geq -k_1 \ln \epsilon$, then

$$\begin{aligned} p(0) &\leq -1 - \frac{\alpha}{\epsilon\theta}\left[1 - \epsilon^{\epsilon\theta k_1}\right] \\ &\to -1 + \alpha k_1 \ln \epsilon = -1 + 7 \ln \epsilon \end{aligned}$$

as $\epsilon \to 0$. Since $p(0) = 6 \ln \epsilon$, this is impossible if ϵ is sufficiently small.

As in §2B, we consider $\Phi(\xi) = p(\xi) + \Lambda^{-1} q(\xi)$.

PROPOSITION 2E.2. *There exists K_2 such that if ϵ is sufficiently small and $0 \leq \xi \leq \xi_\epsilon$, then*

$$\max\{|\Phi(\xi)|, |\Phi'(\xi)|\} \leq K_2 \epsilon^2 (\ln \epsilon)^2 .$$

Proof. We integrate (2B.7b) from ξ to ξ_ϵ to obtain

$$\Phi(\xi) = e^{\epsilon\theta(\xi-\xi_\epsilon)}\Phi(\xi_\epsilon) + \epsilon\theta(\Lambda^{-1}-1)e^{\epsilon\theta\xi}\int_\xi^{\xi_\epsilon} e^{-\epsilon\theta\eta}q(\eta)d\eta\,.$$

We use Proposition 2B.2, Proposition 2E.1, (1.15), and Lemma 2B.3 to conclude that if ϵ is sufficiently small and $0 \le \xi \le \xi_\epsilon$, then

$$\begin{aligned}|\Phi(\xi)| &\le \epsilon^2 K_1 + 2\epsilon^2\theta M\xi_\epsilon \max_{\xi\ge 0}|p(\xi)| \\ &\le \epsilon^2 K_1 + 12\epsilon^2\theta M k_1(\ln\epsilon)^2 \\ &\le \epsilon^2 K_2(\ln\epsilon)^2\end{aligned}$$

for some K_2 which does not depend on ϵ. A similar bound is obtained for $|\Phi'(\xi)|$ if we use (1.15), Lemma 2B.3, and (2B.7a).

The next two corollaries follow from the preceding proposition, (2E.1), and the definitions.

COROLLARY 2E.3. (a) $|T_\xi(\xi) + \Lambda^{-1}Y_\xi(\xi)| \le K_2\epsilon^3(\ln\epsilon)^2$ *for* $0 \le \xi \le \xi_\epsilon$.

(b) $|T_z(0) + \Lambda^{-1}Y_z(0)| \le K_2\epsilon^2(\ln\epsilon)^2$.

COROLLARY 2E.4. *There exist positive constants* α_0 *and* β_0 *such that if* $0 \le \xi \le \xi_\epsilon$, *then* $\alpha_0 < p'(\xi) < \beta_0$, *and* $\alpha_0 < -q'(\xi) < \beta_0$.

3. Proof of Proposition 1.1. Throughout this section we fix the constants ϵ, k, and ℓ. For convenience, we shall drop the subscript ϵ in T_ϵ, Y_ϵ, and θ_ϵ. We set $u' = p$, $v' = q$, and write (1.11) as the first-order system

$$W' = A_\sigma(z)W \tag{3.1}$$

where $W = (u, p, v, q)^T$ and

$$A_\sigma(z) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \sigma + k^2 - \frac{1}{\epsilon^3}Y\phi\left(\frac{T-1}{\epsilon}\right) & \theta & -\frac{1}{\epsilon^2}\phi\left(\frac{T-1}{\epsilon}\right) & 0 \\ 0 & 0 & 0 & 1 \\ \Lambda\frac{1}{\epsilon^3}Y\phi\left(\frac{T-1}{\epsilon}\right) & 0 & \Lambda\sigma + k^2 + \frac{1}{\epsilon^2}\Lambda Y\phi\left(\frac{T-1}{\epsilon}\right) & \Lambda\theta \end{bmatrix}.$$

We say that σ is an eigenvalue if there exists a nontrivial bounded solution of (3.1). It is not hard to prove that σ is an eigenvalue if and only if there exists a nontrivial solution $W(z)$ of (3.1) which satisfies $\lim_{|z|\to\infty} W(z) = O$, the zero vector in $\mathbb{C}^4$, complex four-space. Here we used (1.15). Consider the linear subspaces:

$$\begin{aligned}E_\sigma^- &= \{W(z) : W(z)\text{is a solution of (3.1) and } \lim_{z\to-\infty} \mathrm{W(z)} = \mathrm{O}\}\,, \\ E_\sigma^+ &= \{W(z) : W(z)\text{is a solution of (3.1) and } \lim_{z\to+\infty} \mathrm{W(z)} = \mathrm{O}\}.\end{aligned}$$

Then σ is an eigenvalue if and only if E_σ^- has nontrivial intersection with E_σ^+.

LEMMA 3.1. $\dim \mathrm{E}_\sigma^- = \dim \mathrm{E}_\sigma^+ = 2$.

Proof. Let $A_\sigma^- = \lim_{z\to-\infty} A_\sigma(z)$ and $A_\sigma^+ = \lim_{z\to+\infty} A_\sigma(z)$. Using (1.15), it is a simple manner to show that both A_σ^- and A_σ^+ have two eigenvalues with negative

real parts and two eigenvalues with positive real parts. The result then follows from Coddington and Levinson [2, Thm. 8.1].

Let $\{e_1^-(z), e_2^-(z)\}$ form a basis for E_σ^- and $\{e_1^+(z), e_2^+(z)\}$ form a basis for E_σ^+. Let $\Phi_\sigma(z)$ be the 4×4 matrix whose column vectors are $e_1^-(z)$, $e_2^-(z)$, $e_1^+(z)$, and $e_2^+(z)$. We then define

$$D_\epsilon(\sigma, k, \ell) = \det \Phi_\sigma(0). \tag{3.2}$$

The following two lemmas complete the proof of Proposition 1.1.

LEMMA 3.2. *σ is an eigenvalue if and only if $D_\epsilon(\sigma, k, \ell) = 0$.*

Proof. Recall that σ is an eigenvalue if and only if E_σ^- has nontrivial intersection with E_σ^+. This is true if and only if there exist constants c_1, c_2, c_3, and c_4, not all of which are zero, such that

$$c_1 e_1^- + c_2 e_2^- = c_3 e_1^+ + c_4 e_2^+ .$$

This last statement is true if and only if the columns of $\Phi_\sigma(z)$ are linearly dependent; that is, $\det \Phi_\sigma(z) = 0$ for all z. This is true if and only if $\det \Phi_\sigma(0) = 0$.

LEMMA 3.3. *The vectors e_1^-, e_2^-, e_1^+, and e_2^+ can be chosen so that $D_\epsilon(\sigma, k, \ell)$ is a holomorphic function of σ.*

Proof. See Gardner and Jones [5].

In the next few sections we prove Theorem 1.2 by constructing, to high enough order, vector functions e_1^- and e_2^- in E_σ^-, and e_1^+ and e_2^+ in E_σ^+. We then compute $D_\epsilon(\sigma, k, \ell) = \det \Phi_\sigma(0)$ explicitly. We assume throughout that the translation of the wave is chosen so that (2A.2) holds.

Remarks. (1) Evans [3] defined a function $D(\lambda)$ in his study of nerve impulse equations. That function played a role analogous to the function $D_\epsilon(\sigma, k, \ell)$ used here.

(2) Gardner and Jones [5] have demonstrated how to construct $D(\lambda)$ for more general reaction-diffusion systems. I am grateful to both of them for discussing their work with me.

4. The basic formulas. We fix ϵ, ℓ, k, and σ as in the preceding sections. As before, we drop the subscript ϵ. In this section we derive formulas for linearly independent solutions in E_σ^- and E_σ^+. The basic idea of how to derive the formulas is straightforward, but the actual proofs are quite technical. We shall present the proofs in §6.

A. E_σ^-. We now compute two linearly independent solutions of (3.1) which vanish at $z = -\infty$. The basic ideas is that the reaction terms in (3.1) are very small for $z \le 0$.

Recall that the translation of the wave is chosen so that $T(0) = 1 + 6\epsilon \ln \epsilon$. Since T and ϕ are monotone increasing functions, it follows that if $z \le 0$, then

$$\phi\left(\frac{T(z) - 1}{\epsilon}\right) \le \phi\left(\frac{T(0) - 1}{\epsilon}\right) = \epsilon^6 .$$

Hence, the nonlinear terms in (1.11) are very small. This leads us to consider (1.11) without the reaction terms; that is

$$\begin{aligned} u'' - \theta u' &= (\sigma + k^2)u, \\ \Lambda^{-1} v'' - \theta v' &= (\sigma + \Lambda^{-1} k^2) v. \end{aligned} \tag{4A.1}$$

Two linearly independent solutions of (4A.1) which decay at $z = -\infty$ are:

$$(u_1(z), v_1(z)) = (e^{r_1 z}, 0) \quad \text{and} \quad (u_2(z), v_2(z)) = (0, e^{r_2 z})$$

where

$$(4A.2)\quad r_1 = \frac{\theta + \sqrt{\theta^2 + 4(\sigma + k^2)}}{2}, \qquad r_2 = \frac{\theta + \sqrt{\theta^2 + 4\Lambda^{-1}(\sigma + \Lambda^{-1}k^2)}}{2\Lambda^{-1}}.$$

Hence, our guesses for $e_1^-(z)$ and $e_2^-(z)$ are

$$(4A.3)\quad \hat{e}_1^-(z) = \begin{pmatrix} u_1(z) \\ u_1'(z) \\ v_1(z) \\ v_1'(z) \end{pmatrix} = \begin{pmatrix} e^{r_1 z} \\ r_1 e^{r_1 z} \\ 0 \\ 0 \end{pmatrix}, \qquad \hat{e}_2^-(z) = \begin{pmatrix} u_2(z) \\ u_2'(z) \\ v_2(z) \\ v_2'(z) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ e^{r_2 z} \\ r_2 e^{r_2 z} \end{pmatrix}.$$

In §6 we shall prove that these are indeed good guesses.

PROPOSITION 4A.1. *There exists ϵ_0 such that if $0 < \epsilon \le \epsilon_0$, then there exist vector functions $e_1^-(z)$ and $e_2^-(z)$ in E_σ^- such that for $z \le 0$,*

$$(4A.4)\qquad \|e_1^-(z) - \hat{e}_1^-(z)\| \le \epsilon^2 \quad \text{and} \quad \|e_2^-(z) - \hat{e}_2^-(z)\| \le \epsilon^2.$$

B. e_1^+. We now construct, to high enough order, a trajectory $e_1^+(z)$ in E_σ^+. As in §2B, it is convenient to work in the variables $p = T - 1/\epsilon$, $q = Y/\epsilon$, and $\xi = z/\epsilon$. In these new variables, (1.11) becomes

$$(4B.1)\qquad \begin{aligned} &u'' - \epsilon\theta u' + q\phi(p)u + \phi(p)v = \epsilon^2(\sigma + k^2)u, \\ &\Lambda^{-1}u'' - \epsilon\theta v' - q\phi(p)u - \phi(p)v = \epsilon^2(\sigma + \Lambda^{-1}k^2)v \end{aligned}$$

where differentiation is with respect to ξ.

The basic idea of this section is that the right-hand side of (4B.1) is small compared to the left-hand side. Hence, we look for a solution of (4B.1) of the form:

$$(4B.2)\qquad u(\xi) = u_1(\xi) + \epsilon^2 u_2(\xi) \quad \text{and} \quad v(\xi) = v_1(\xi) + \epsilon^2 v_2(\xi)$$

where $(u_1(\xi), v_1(\xi))$ satisfies the system:

$$(4B.3)\qquad \begin{aligned} &u_1'' - \epsilon\theta u_1' + q\phi(p)u_1 + \phi(p)v_1 = 0, \\ &\Lambda^{-1}v_1'' - \epsilon\theta v_1' - q\phi(p)u_1 - \phi(p)v_1 = 0. \end{aligned}$$

Of course, we must also impose the boundary conditions:

$$(4B.4)\qquad \lim_{\xi\to\infty}(u_1(\xi), v_1(\xi)) = \lim_{\xi\to\infty}(u_2(\xi), v_2(\xi)) = (0,0).$$

It is not hard to check that for $(u_1(\xi), v_1(\xi))$ we may take

$$(4B.5)\qquad u_1(\xi) = p'(\xi) \quad \text{and} \quad v_1(\xi) = q'(\xi).$$

That is, e_1^+ will be a perturbation of the derivative of the wave.

We can now make a preliminary guess for e_1^+. When we compute e_1^+, we must be careful to distinguish between the variables z and ξ. Of course, we are only really interested in $e_1^+(0)$. Our guess for $e_1^+(0)$ is then

$$\text{(4B.6)}\quad \hat{e}_1^+(0) = \begin{pmatrix} u(0) \\ u_z(0) \\ v(0) \\ v_z(0) \end{pmatrix} = \begin{pmatrix} p_\xi(0) + \epsilon^2 u_2(0) \\ \frac{1}{\epsilon} p_{\xi\xi}(0) + \epsilon u_{2\xi}(0) \\ q_\xi(0) + \epsilon^2 v_2(0) \\ \frac{1}{\epsilon} q_{\xi\xi}(0) + \epsilon v_{2\xi}(0) \end{pmatrix} = \begin{pmatrix} T_z(0) + \epsilon^2 u_2(0) \\ T_{zz}(0) + \epsilon u_{2\xi}(0) \\ Y_z(0) + \epsilon^2 v_2(0) \\ Y_{zz}(0) + \epsilon v_{2\xi}(0) \end{pmatrix}.$$

We now use the results of §2 to express $T_{zz}(0)$, $Y_z(0)$, and $Y_{zz}(0)$ in terms of $T_z(0)$. From (2D.3), and Corollary 2E.3, we have that

$$\text{(4B.7)}\quad \begin{aligned} &\text{(a)}\quad T_{zz}(0) = \theta T_z(0) + \text{h.o.t.} \\ &\text{(b)}\quad Y_z(0) = -\Lambda T_z(0) + \text{h.o.t.} \\ &\text{(c)}\quad Y_{zz}(0) = \theta\Lambda Y_z(0) + \text{h.o.t.} = -\theta\Lambda^2 \mathrm{T}_z(0) + \text{h.o.t.} \end{aligned}$$

By "h.o.t." we mean, throughout this paper, a term which is $0(\epsilon^2 \ln^4 \epsilon)$. Therefore, (4B.6) becomes,

$$\text{(4B.8)}\quad \hat{e}_1^+(0) = \begin{pmatrix} T_z(0) + \epsilon^2 u_2(0) \\ \theta T_z(0) + \epsilon u_{2\xi}(0) + \text{h.o.t.} \\ -\Lambda T_z(0) + \epsilon^2 v_2(0) + \text{h.o.t.} \\ -\theta\Lambda^2 T_z(0) + \epsilon v_{2\xi}(0) + \text{h.o.t.} \end{pmatrix}.$$

We must now worry about the terms $u_2(0)$, $u_{2\xi}(0)$, $v_2(0)$, and $v_{2\xi}(0)$. In §6 we prove the following proposition.

PROPOSITION 4B.1. *There exists a positive constant C such that if ϵ is sufficiently small, then*

$$\text{(4B.9)}\quad \begin{aligned} &\text{(a)}\quad |u_2(0)| \le C(\ln\epsilon)^2, \qquad |u_{2\xi}(0)| \le C(\ln\epsilon)^2 \\ &\text{(b)}\quad |v_2(0)| \le C(\ln\epsilon)^2 \\ &\text{(c)}\quad |u_{2\xi}(0) + \Lambda^{-1} v_{2\xi}(0)| \le \epsilon C(\ln\epsilon)^2. \end{aligned}$$

It then follows that

$$\text{(4B.10)}\quad \hat{\epsilon}_1^+(0) = \begin{pmatrix} T_z(0) + \text{h.o.t.} \\ \theta T_z(0) + \epsilon u_{2\xi}(0) + \text{h.o.t.} \\ -\Lambda T_z(0) + \text{h.o.t.} \\ -\theta\Lambda^2 T_z(0) - \epsilon\Lambda u_{2\xi}(0) + \text{h.o.t.} \end{pmatrix}.$$

Finally, note that by (2D.2), $T_z(0)$ is bounded away from zero. Hence, we may divide each term in (4B.10) by $T_z(0)$, and set

$$\text{(4B.11)}\quad P_1 = \frac{u_{2\xi}(0)}{T_z(0)},$$

to obtain

$$e_1^+(0) = \begin{pmatrix} 1 \\ \theta + \epsilon P_1 \\ -\Lambda \\ -\theta\Lambda^2 - \epsilon P_1 \end{pmatrix} + \text{h.o.t.} \tag{4B.12}$$

Note that we have not computed P_1 explicitly. When we evaluate $D_\epsilon(\sigma, k, \ell)$ by taking a certain determinant, this term will cancel, to high enough order.

C. e_2^+. In this section we construct, to high enough order, another trajectory in E_σ^+. We begin by rewriting (4B.1) as

$$\begin{aligned} &u'' - \epsilon\theta u' + q\phi(p)u + \phi(p)v - \epsilon^2(\sigma + k^2)u = 0, \\ &\Lambda^{-1}v'' - \epsilon\theta\Lambda^{-1}v' - q\phi(p)u - \phi(p)v - \epsilon^2(\sigma + k^2)\Lambda^{-1}v \\ &\qquad = \epsilon\theta(1 - \Lambda^{-1})v' + \epsilon^2(1 - \Lambda^{-1})\sigma v. \end{aligned} \tag{4C.1}$$

Let $w(\xi) = u(\xi) + \Lambda^{-1}v(\xi)$. Adding the two equations in (4C.1), we find that $w(\xi)$ satisfies the equation

$$w'' - \epsilon\theta w' - \epsilon^2(\sigma + k^2)w = \epsilon\theta(1 - \Lambda^{-1})v' + \epsilon^2(1 - \Lambda^{-1})\sigma v\,. \tag{4C.2}$$

Since the right-hand side of (4C.2) is of order ϵ^2, we guess that $w(\xi) = w_1(\xi)+$ h.o.t. where $w_1(\xi)$ satisfies the equation

$$w_1'' - \epsilon\theta w_1' - \epsilon^2(\sigma + k^2)w_1 = 0\,. \tag{4C.3}$$

For $w_1(\xi)$ we take

$$w_1(\xi) = \epsilon e^{\epsilon\gamma\xi} \quad \text{where } \gamma = \frac{\theta - \sqrt{\theta^2 + 4(\sigma + k^2)}}{2}\,. \tag{4C.4}$$

Recall from (1.15) that $\operatorname{Re}\gamma < 0$. Choose P_2 and P_3 so that

$$u(0) = P_2 \quad \text{and} \quad u_\xi(0) = \epsilon P_3. \tag{4C.5}$$

From the definition of $w(\xi)$, we expect that

$$\begin{aligned} v(0) &= -\Lambda u(0) + \Lambda w(0) = -\Lambda P_2 + \epsilon\Lambda + \text{h.o.t.}\,, \\ v_\xi(0) &= -\Lambda u'(0) + \Lambda w'(0) = -\epsilon\Lambda P_3 + \epsilon^2\Lambda\gamma + \text{h.o.t.} \end{aligned} \tag{4C.6}$$

Therefore,

$$u_z(0) = \frac{1}{\epsilon}u_\xi(0) = P_3 \quad \text{and} \quad \mathrm{v_z}(0) = \frac{1}{\epsilon}\mathrm{v}_\xi(0) = -\Lambda\mathrm{P}_3 + \epsilon\Lambda\gamma + \text{h.o.t.} \tag{4C.7}$$

Our guess for $e_2^+(0)$ is then

$$e_2^+(0) = \begin{pmatrix} u(0) \\ u_z(0) \\ v(0) \\ v_z(0) \end{pmatrix} = \begin{pmatrix} P_2 \\ P_3 \\ -\Lambda P_2 + \epsilon\Lambda \\ -\Lambda P_3 + \epsilon\Lambda\gamma \end{pmatrix} + \text{h.o.t.} \tag{4C.8}$$

In §9 we prove the following proposition.

PROPOSITION 4C.1. *For ϵ sufficiently small, there exists a solution $e_2^+(\xi)$ in E_σ^+ such that $e_2^+(0)$ is given by (4C.8) for some constants P_2 and P_3. Moreover, $|P_3 - \frac{1}{\theta}| \le c_1\epsilon|\ln\epsilon|^4$ and $|P_2| \le c_1\epsilon(\ln\epsilon)^2$, for some constant c_1 which does not depend on ϵ.*

Note that we do not compute P_2 to highest order. This is because when we compute $D_\epsilon(\sigma, k, \ell)$ by computing a certain determinant, P_2 will cancel to high enough order.

5. Computation of $D_\epsilon(\sigma, k, \ell)$. We are now ready to compute $D_\epsilon(\sigma, k, \ell)$. Actually, we first compute another complex valued function $\hat{D}_\epsilon(\sigma, k, \ell)$ which satisfies Proposition 1.1. Then $D_\epsilon(\sigma, k, \ell)$ will be equal to some constant multiple of $\hat{D}_\epsilon(\sigma, k, \ell)$. The constant will be chosen so that Theorem 1.2 is satisfied.

From (3.2), Proposition 4A.1, (4B.12), and (4C.8), we have that

$$\hat{D}_\epsilon(\sigma, k, \ell) = \det \begin{bmatrix} 1 & 0 & 1 & P_2 \\ r_1 & 0 & \theta + \epsilon P_1 & P_3 \\ 0 & 1 & -\Lambda & -\Lambda P_2 + \epsilon\Lambda \\ 0 & r_2 & -\theta\Lambda^2 - \epsilon P_1 & -\Lambda P_3 + \epsilon\Lambda\gamma \end{bmatrix} + \text{h.o.t.}$$

A straightforward computation shows that

$$\begin{aligned} &\hat{D}_\epsilon(\sigma, k, \ell) \\ &\quad = \det \begin{bmatrix} 1 & 0 & 1 & P_2 \\ 0 & 0 & \theta + \epsilon P_1 - r_1 & P_3 - r_1 P_2 \\ 0 & 1 & -\Lambda & -\Lambda P_2 + \epsilon\Lambda \\ 0 & 0 & -\theta\Lambda^2 - \epsilon P_1 + r_2\Lambda & -\Lambda(r_2 P_2 - P_3) + \epsilon\Lambda(\gamma - r_2) \end{bmatrix} + \text{h.o.t.} \\ &\quad = \Lambda(\theta + \epsilon P_1 - r_1)(r_2 P_2 - P_3 + \epsilon\gamma - \epsilon r_2) \\ &\qquad - (P_3 - r_1 P_2)(-\theta\Lambda^2 - \epsilon P_1 + r_2\Lambda) + \text{h.o.t.} \\ &\quad = \Lambda P_3(r_1 - r_2) + \theta P_3\Lambda(\Lambda - 1) + \epsilon\Lambda(\theta - r_1)(\gamma - r_2) \\ &\qquad + \theta\Lambda P_2(r_2 - \Lambda r_1) + \epsilon P_1 P_3(1 - \Lambda) + \epsilon P_1 P_2(r_2\Lambda - r_1) \\ &\qquad + \epsilon^2\Lambda P_1(\gamma - r_2) + \text{h.o.t.} \end{aligned} \tag{5.1}$$

Let $\Gamma_\epsilon \equiv \sqrt{(\theta^2/4) + \sigma + k^2}$. Recall that we are not writing the subscript ϵ on θ_ϵ. A straightforward computation shows that

$$r_1 = \frac{\theta}{2} + \Gamma_\epsilon, \qquad \gamma = \frac{\theta}{2} - \Gamma_\epsilon,$$

(5.2)

$$r_2 = r_1 - \epsilon\ell\left[\frac{\sigma + 2k^2}{2\Gamma_\epsilon} - \frac{\theta}{2} - \Gamma_\epsilon\right] + \text{h.o.t.}$$

We combine (5.1), (5.2), (1.15), Proposition 4C.1, and Proposition 4B.1 to conclude that

$$\begin{aligned}
\hat{D}_\epsilon(\sigma, k, \ell) =& \Lambda P_3(r_1 - r_2) + \epsilon\ell\Lambda\theta P_3 + \epsilon\Lambda(\theta - r_1)(\gamma - r_2) + \text{h.o.t.} \\
=& \left(\frac{\epsilon\ell}{\theta}\right)\left(\frac{\sigma + 2k^2}{2\Gamma_\epsilon} - \frac{\theta}{2} - \Gamma_\epsilon\right) + \epsilon\ell - \epsilon\left(\frac{\theta}{2} - \Gamma_\epsilon\right)(2\Gamma_\epsilon) + \text{h.o.t.} \\
=& \left(\frac{\epsilon}{2\Gamma_\epsilon\theta}\right)\ell(\sigma + 2k^2 + \theta\Gamma_\epsilon - 2\Gamma_\epsilon^2) - \left(\frac{\epsilon}{2\Gamma_\epsilon\theta}\right)4\theta\Gamma_\epsilon^2\left(\frac{\theta}{2} - \Gamma_\epsilon\right) + \text{h.o.t.} \\
=& \left(\frac{\epsilon}{2\Gamma_\epsilon\theta}\right)\left\{\ell\left(\sigma + 2k^2 + \theta\Gamma_\epsilon - \frac{\theta^2}{2} - 2\sigma - 2k^2\right)\right. \\
& \qquad \left. -4\theta\Gamma_\epsilon^2\left(\frac{\theta}{2} - \Gamma_\epsilon\right)\right\} + \text{h.o.t.} \\
=& \left(\frac{\epsilon}{2\Gamma_\epsilon\theta}\right)\left\{\ell\left(\theta\Gamma_\epsilon - \frac{\theta^2}{2} - \sigma\right) - 4\theta\Gamma_\epsilon^2\left(\frac{\theta}{2} - \Gamma_\epsilon\right)\right\} + \text{h.o.t.}
\end{aligned}$$

Let $D_\epsilon(\sigma, k, \ell) = -(2\Gamma_\epsilon\theta/\epsilon)\,\hat{D}_\epsilon(\sigma, k, \ell)$. Then,

$$D_\epsilon(\sigma, k, \ell) = -\ell\left(\theta\Gamma_\epsilon - \frac{\theta^2}{2} - \sigma\right) - 4\theta\Gamma_\epsilon^2\left(\frac{\theta}{2} - \Gamma_\epsilon\right) + h(\epsilon)$$

where $\lim_{\epsilon\to 0} h(\epsilon) = 0$. To complete the proof of Theorem 1.2, we recall that $\lim_{\epsilon\to 0}\theta_\epsilon = \sqrt{2}$. Hence,

$$\lim_{\epsilon\to 0}\Gamma_\epsilon = \sqrt{\frac{1}{2} + \sigma + k^2} = \frac{1}{\sqrt{2}}\Gamma$$

where Γ is given by (1.13). This implies that

$$D_\epsilon(\sigma, k, \ell) = \ell(\Gamma - 1 - \sigma) - 2\Gamma^2(\Gamma - 1) + h_1(\epsilon)$$

where $\lim_{\epsilon\to 0} h_1(\epsilon) = 0$.

6. Proofs of the basic formulas.

A. Proof of Proposition 4A.1. Throughout this section we use the notation of §4A.1. The proof of Proposition 4A.1 is straightforward so we shall only outline the details. In later sections, similar, but more complicated, arguments will be worked out in detail. We only outline the proof of the first inequality in (4A.4), because the proof of the second one is almost identical.

If $z < z^\epsilon$, then $\phi(T(z) - 1/\epsilon) = 0$. Hence, $e_1^-(z) \equiv \hat{e}_1^-(z) \in E_\sigma^-$ for $z < z^\epsilon$. Throughout this proof we let $(u(z), v(z))$ be the solution of (1.11) such that $(u(z), v(z)) = (u_1(z), v_1(z))$ for $z \le z^\epsilon$.

Let $\beta(z) = u(z) - u_1(z)$ for $z^\epsilon \le z \le 0$. Then $(\beta(z), v(z))$ satisfies the system

$$\beta'' - \epsilon\theta\beta' - (\sigma + k^2)\beta = -\frac{1}{\epsilon^3} Y\phi\left(\frac{T-1}{\epsilon}\right)(\beta + u_1) - \frac{1}{\epsilon^2}\phi\left(\frac{T-1}{\epsilon}\right) v,$$

(6A.1)

$$\Lambda^{-1}v'' - \epsilon\theta v' - (\sigma + \Lambda^{-1}k^2)v = \frac{1}{\epsilon^3} Y\phi\left(\frac{T-1}{\epsilon}\right)(\beta + u_1) + \frac{1}{\epsilon^2}\phi\left(\frac{T-1}{\epsilon}\right) v$$

together with the boundary conditions

$$(\beta(z^\epsilon), v(z^\epsilon)) = (0,0), \quad \beta'(z^\epsilon+) = -\left(Y(z^\epsilon)u_1(z^\epsilon)\exp\left(\frac{T_* - 1}{\epsilon}\right)\right)/(\epsilon^3 T'(z^\epsilon)),$$

(6A.2)

$$v'(z^\epsilon+) = \left(\Lambda Y(z^\epsilon)u_1(z^\epsilon)\exp\left(\frac{T_* - 1}{\epsilon}\right)\right)/(\epsilon^3 T'(z^\epsilon)).$$

Here we used the jump conditions (1.12). In §2C we showed that $T'(z^\epsilon) = \theta T_*$. It follows that if ϵ is sufficiently small, then

$$|\beta'(z^\epsilon+)| < \epsilon^4 \quad \text{and} \quad |v'(z^\epsilon+)| < \epsilon^4. \tag{6A.3}$$

We use an iteration scheme to prove the existence of a solution of (6A.1), (6A.2). The existence proof will also give us the desired estimates. To set up the iteration scheme, we write (6A.1), for $j = 1, 2, \cdots$, as

$$\begin{aligned} \beta_j'' - \theta\beta_j' - (\sigma + k^2)\beta_j &= -h_{j-1}(z), \\ \Lambda^{-1}v_j'' - \theta v_j' - (\sigma + \Lambda^{-1}k^2)v_j &= h_{j-1}(z) \end{aligned} \tag{6A.4}$$

where $h_{j-1}(z) = (1/\epsilon^3)Y\phi(T - 1/\epsilon)(\beta_{j-1} + u_1) + (1/\epsilon^2)\phi(T - 1/\epsilon)v_{j-1}$. We assume that $(\beta_0, v_0) \equiv (0,0)$, and for $j \ge 1$, (β_j, v_j) satisfies the boundary conditions (6A.2). Let

$$\mathcal{S} = \{(\phi(z), \psi(z)) : \max(|\phi(z)|, |\phi'(z)|, |\psi(z)|, |\psi'(x)|) \le \epsilon^2 \text{ for } z^\epsilon \le z \le 0\}.$$

Using induction we prove that each $(\beta_j, v_j) \in \mathcal{S}$. Clearly, $(\beta_0, v_0) \in \mathcal{S}$. Suppose that $(\beta_{j-1}, v_{j-1}) \in \mathcal{S}$. We shall only obtain the necessary estimates for $\beta_j(z)$ since the estimates for $v_j(z)$ are obtained in a similar fashion.

From the variation of constants formula it follows that

$$\beta_j(z) = e^{r_1(z-z^\epsilon)}\Psi_1(z) + e^{m_1(z-z^\epsilon)}\Psi_2(z) \tag{6A.5}$$

where

$$\text{(a)} \quad r_1 = \frac{\theta + \sqrt{\theta^2 + 4(\sigma + k^2)}}{2}, \qquad m_1 = \frac{\theta - \sqrt{\theta^2 + 4(\sigma + k^2)}}{2},$$

$$\text{(b)} \quad \Psi_1(z) = \frac{1}{\sqrt{\theta^2 + 4(\sigma + k^2)}} \int_{z_\epsilon}^{z} e^{(m_1 - \theta)(\eta - z^\epsilon)} h_{j-1}(\eta)\, d\eta + \lambda,$$

(6A.6)

$$\text{(c)} \quad \Psi_2(z) = -\frac{1}{\sqrt{\theta^2 + 4(\sigma + k^2)}} \int_{z_\epsilon}^{z} e^{(r_1 - \theta)(\eta - z^\epsilon)} h_{j-1}(\eta)\, d\eta - \lambda,$$

$$\text{(d)} \quad \lambda = \beta'(z^\epsilon)/\sqrt{\theta^2 + 4(\sigma + k^2)}.$$

Because $\phi(T-1/\epsilon) \leq \epsilon^6$ for $z \leq 0$, and $(\beta_{j-1}, v_{j-1}) \in \mathcal{S}$, it follows that $|h_{j-1}(z)| \leq 2\epsilon^3$. From Proposition 2D.1, $|z^\epsilon|$ is bounded, independent of ϵ. It now follows from (6A.4) and (6A.6) that there exists a constant C, which does not depend on ϵ, such that if $z^\epsilon \leq z \leq 0$, and ϵ is sufficiently small, then $|\Psi_1(z)| \leq C\epsilon^3$ and $|\Psi_2(z)| \leq C\epsilon^3$. It then follows from (6A.5) and Proposition 2D.1, that if $z^\epsilon \leq z \leq 0$, and ϵ is sufficiently small, then $|\beta_j(z)| \leq \epsilon^2$. Moreover, since

$$\beta_j'(z) = r_1 e^{r_1(z-z^\epsilon)}\Psi_1(z) + m_1 e^{m_1(z-z^\epsilon)}\Psi_2(z),$$

a similar argument shows that if $z^\epsilon \leq z \leq 0$, and ϵ is sufficiently small, then $|\beta_j'(z)| \leq \epsilon^2$. In a similar fashion we may obtain the necessary estimates for $|v_j(z)|$ and $|v_j'(z)|$. Hence, $(\beta_j, v_j) \in \mathcal{S}$. In order to complete the proof of the proposition, we must show that a subsequence of $\{(\beta_j, v_j)\}$ converges to functions $(\beta(z), v(z))$ in $\mathcal{S}$ such that $(\beta(z), v(z))$ is a solution of (6A.1), (6A.2). This is a straightforward argument, especially with the estimates we have obtained so far. The complete details for this type of argument will be given in the next sections when we consider E_σ^+.

B. Solutions of the homogeneous equations. We now prove a very important preliminary result. Consider the homogeneous equation

$$\Gamma'' - \epsilon\theta\Gamma' + [(q-\Lambda)\phi + \epsilon^2\lambda]\Gamma = 0 \tag{6B.1}$$

where $\xi, q(\xi), p(\xi)$, and $\varphi(p)$ are as in earlier sections, and λ is a complex number with $|\lambda| \leq \lambda_0$. Let

$$\delta_1 = \frac{\epsilon\theta - \sqrt{4\Lambda + \epsilon^2\theta^2 - \epsilon^2\lambda}}{2} \quad \text{and} \quad \delta_2 = \frac{\epsilon\theta + \sqrt{4\Lambda + \epsilon^2\theta^2 - \epsilon^2\lambda}}{2}$$

be the characteristic roots of (6B.1) at $\xi = \infty$. We assume that ϵ is so small that $\operatorname{Re}\delta_1 < 0 < \operatorname{Re}\delta_2$.

PROPOSITION 6B.1. *There exist positive constants M^1, M_1, C_1, and ρ, and solutions $\Gamma_1(\xi)$ and $\Gamma_2(\xi)$ of (6B.1) such that for ϵ sufficiently small,*

$$\begin{aligned}
&\text{(a)} \quad \max\{|\Gamma_i(\xi)|, |\Gamma_i'(\xi)|\} \leq C_1 e^{\delta_i(\xi-\xi_\epsilon)} \quad \text{for } \xi \geq \xi_\epsilon, \qquad i = 1,2,\\
&\text{(b)} \quad M^1 < |\Gamma_2'(\xi)| < M_1 \text{ and } |\Gamma_2(\xi)| \leq M_1(|\ln\epsilon| + 1) \quad \text{for } 0 \leq \xi \leq \xi_\epsilon,\\
&\text{(c)} \quad \Gamma_1(\xi_\epsilon) = p'(\xi_\epsilon), \qquad -\rho \leq \operatorname{Re}\Gamma_1'(\xi_\epsilon) \leq 0,\\
&\text{(d)} \quad 1 \leq \operatorname{Re}\Gamma_2(\xi_\epsilon) \leq \rho, \qquad 1 \leq \operatorname{Re}\Gamma_2'(\xi_\epsilon) \leq \rho.
\end{aligned} \tag{6B.2}$$

Proof. From Coddington and Levinson [2], there exist solutions $\Gamma_1(\xi)$ and $\hat{\Gamma}_2(\xi)$ of (6B.1) and a positive constant C_1 such that for $\xi \geq \xi_\epsilon$,

$$\max\{|\Gamma_1(\xi)|, |\Gamma_1'(\xi)|\} \leq C_1 e^{\delta_1(\xi-\xi_\epsilon)}, \qquad \max\{|\hat{\Gamma}_2(\xi)|, |\hat{\Gamma}_2'(\xi)|\} \leq C_1 e^{\delta_2(\xi-\xi_\epsilon)}.$$

It is not hard to see that C_1 can be chosen independently of ϵ and $\Gamma_1(\xi)$ can be chosen so that if $\rho > C_1$, then (6B.2c) is satisfied. Actually much stronger properties of Γ_1 are proven in Proposition 6C.2.

We now consider $\Gamma_2(\xi)$. Choose ζ_ϵ such that $p(\zeta_\epsilon) = -P_0$ where $P_0 > 1$ is to be chosen shortly. Since $p(\xi_\epsilon) = -1$ and $p'(\xi) > 0$ for all ξ, it follows that $\zeta_\epsilon < \xi_\epsilon$. Moreover, from Corollary 2E.4, we conclude that $(\xi_\epsilon - \zeta_\epsilon)$ is bounded from above and below, independently of ϵ. Let $\Gamma_2(\xi)$ be the solution of (6B.1) such that $\Gamma_2(\zeta_\epsilon) = 0$ and $\Gamma_2'(\zeta_\epsilon) = P_1$ where $P_1 > 0$ is determined as follows. Because $(\xi_\epsilon - \zeta_\epsilon)$ is bounded

from above and below, and $\Gamma_2'(\zeta_\epsilon) = P_1 > 0$, it is clear that we can choose P_1 and ρ; independent of ϵ, such that $1 \le \operatorname{Re}\Gamma_2(\xi_\epsilon) \le \rho$ and $1 \le \operatorname{Re}\Gamma_2'(\xi_\epsilon) \le \rho$. This verifies (6B.2d). Moreover, because $\Gamma_2(\xi)$ can be written as a linear combination of $\Gamma_1(\xi)$ and $\hat{\Gamma}_2(\xi)$, (6B.2a) follows. It remains to show that P_0 can be chosen so that (6B.2b) holds.

We first need some preliminary estimates. It follows from Corollary 2E.4 and Proposition 2E.2 that if $0 < \xi < \zeta_\epsilon$, then

$$p(\xi) \le p(\zeta_\epsilon) + \alpha_0(\xi - \zeta_\epsilon) = -P_0 + \alpha_0(\xi - \zeta_\epsilon),$$
$$q(\xi) \le q(\zeta_\epsilon) - \beta_0(\xi - \zeta_\epsilon) \le 2P_0 - \beta_0(\xi - \zeta_\epsilon).$$

Therefore,

$$(q - \Lambda)\varphi(P) \le (2P_0 + \beta_0(\zeta_\epsilon - \xi) - \Lambda)\exp(-P_0 + \alpha_0(\xi - \zeta_\epsilon)). \tag{6B.3}$$

Now let $m(\xi) = \Gamma_2(\xi)/\Gamma_2'(\xi)$. Then $m' + \epsilon\theta m = 1 + [(q-\Lambda)\varphi(p) + \epsilon^2\lambda]m^2$. We multiply both sides of this equation by $e^{\epsilon\theta(\xi-\zeta_\epsilon)}$ and let $X(\xi) = e^{\epsilon\theta(\xi-\zeta_\epsilon)}m(\xi)$ to obtain

$$X'(\xi) = e^{\epsilon\theta(\xi-\zeta_\epsilon)} + [(q - \Lambda)\varphi(p) + \epsilon^2\lambda]e^{\epsilon\theta(\xi-\zeta_\epsilon)}X^2. \tag{6B.4}$$

Therefore, from (6B.3),

$$|X'(\xi)| \le e^{\epsilon\theta(\xi-\zeta_\epsilon)} + \{[2P_0 + \beta_0(\zeta_\epsilon - \xi) - \Lambda]e^{-P_0}e^{\alpha_0(\xi-\zeta_\epsilon)} + \epsilon^2\lambda_0\}e^{-\epsilon\theta(\xi-\zeta_\epsilon)}|X|^2. \tag{6B.5}$$

Note that $X(\zeta_\epsilon) = 0$ and, from (6B.4), $X'(\zeta_\epsilon) = 1$. Hence, there exist $\delta > 0$ such that $|X(\xi)| < 2(\zeta_\epsilon - \xi)$ for $\zeta_\epsilon - \delta < \xi < \zeta_\epsilon$. We claim that

$$|X(\xi)| < 2(\zeta_\epsilon - \xi) \quad \text{for } 0 \le \xi < \zeta_\epsilon. \tag{6B.6}$$

Suppose that (6B.6) is not true and $\xi_0 = \sup\{\xi < \zeta_\epsilon : |X(\xi)| = 2(\zeta_\epsilon - \xi)\}$. Since $|X(\xi)| \le 2(\zeta_\epsilon - \xi)$ for $\xi_0 \le \xi \le \zeta_\epsilon$, we conclude from (6B.5) that if $P_0 > \Lambda$, ϵ is sufficiently small, and $\xi_0 \le \xi \le \zeta_\epsilon$, then

$$|X'(\xi)| \le 1 + \{[2P_0 + \beta_0(\zeta_\epsilon - \xi)]e^{-P_0}e^{\alpha_0(\xi-\zeta_\epsilon)} + \epsilon^2\lambda_0\}e^{\epsilon\theta(\zeta_\epsilon-\xi)}4(\zeta_\epsilon - \xi)^2. \tag{6B.7}$$

A straightforward computation shows that if P_0 is sufficiently large, and $\xi < \zeta_\epsilon$, then

$$[2P_0 + \beta_0(\zeta_\epsilon - \xi)]e^{-P_0}e^{\alpha_0(\xi-\zeta_\epsilon)}4(\zeta_\epsilon - \epsilon)^2 < \tfrac{1}{2}. \tag{6B.8}$$

Together with (6B.7) and Proposition 2E.1, this implies that if ϵ is sufficiently small and $\xi_0 \le \xi \le \zeta_\epsilon$, then

$$\begin{aligned} |X'(\xi)| &\le 1 + \frac{1}{2} + \epsilon^2\lambda_0 8\xi_\epsilon^2 \\ &\le 1 + \frac{1}{2} + 8\epsilon^2\lambda_0 k_1^2(\ln\epsilon)^2 \\ &< 2\,. \end{aligned}$$

This however implies that $|X(\xi_0)| < 2(\zeta_\epsilon - \xi_0)$ which contradicts the definition of ξ_0.

We have now verified (6B.6). Using the definitions we now conclude that if $0 \le \xi \le \zeta_\epsilon$, and ϵ is sufficiently small, then

$$|\Gamma_2(\xi)/\Gamma_2'(\xi)| \equiv |m(\xi)| = |\Gamma(\xi)|e^{\epsilon\theta(\zeta_\epsilon - \xi)} \le 2(\zeta_\epsilon - \xi)e^{\epsilon\theta(\zeta_\epsilon - \xi)} < 4(\zeta_\epsilon - \xi).$$

Together with (6B.1) and (6B.3) this implies that for $0 \le \xi \le \xi_\epsilon$ and ϵ sufficiently small,

$$|\Gamma_2''(\xi)| = |\epsilon\theta\Gamma_2'(\xi) + [(q-\Lambda)\phi(p) + \epsilon^2\lambda]\Gamma_2(\xi)| \le \epsilon\theta|\Gamma_2'(\xi)| + \{[2P_0 + \beta_0(\zeta_\epsilon - \xi)]e^{-P_0}e^{\alpha_0(\xi-\zeta_\epsilon)} + \epsilon^2\lambda_0\}4(\zeta_\epsilon - \xi)|\Gamma_2'(\xi)| .$$

This is a separable equation with the dominant term being $e^{-P_0}e^{\alpha(\xi-\zeta_\epsilon)}$. We integrate this equation and find that, since $\alpha(\xi - \zeta_\epsilon) < 0$, there exists $M^1 > 0$ such that if P_0 is sufficiently large, then $|\Gamma_2'(\xi)| > M^1$ for $0 \le \xi \le \xi_\epsilon$. This is the desired lower bound for $|\Gamma_2'(\xi)|$ needed for (6B.2b). The upper bound is actually easier, and is obtained in a similar fashion.

C. The inhomogeneous equation. In this section we prove two more preliminary results. For both results we consider equations of the form

$$u'' - \epsilon\theta u' + [(q-\Lambda)\varphi(p) + \epsilon^2\lambda]u = g(\xi) \tag{6C.1}$$

where $q(\xi), p(\xi)$, and λ are as in the preceding section and $g(\xi)$ is a continuous function which satisfies, for some positive constants M_2 and α_1,

$$|g(\xi)| \le \begin{cases} M_2 & \text{for } 0 \le \xi \le \xi_\epsilon \\ M_2 e^{-\alpha_1(\xi - \xi_\epsilon)} & \text{for } \xi \le \xi_\epsilon. \end{cases} \tag{6C.2}$$

Here, M_2 may depend on ϵ, but α_1, does not. Let δ_1 and δ_2 be as in the previous section. We assume that

$$\epsilon\theta < \tfrac{1}{2}\mathrm{Re}\,\delta_2 \quad \text{and} \quad \alpha_1 < \tfrac{1}{2}\mathrm{Re}\,(\delta_2 - \epsilon\theta). \tag{6C.3}$$

PROPOSITION 6C.1. *There exists a solution $u(\xi)$ of (6C.1) such that $\lim_{\xi\to\infty} u(\xi) = 0$, and, for some positive constant M_3 which does not depend on ϵ or α_1,*

$$\max\{|u(\xi)|, \quad |u'(\xi)|\} \le \begin{cases} M_2M_3|\ln\epsilon|(\xi_\epsilon - \xi + 1) & \text{for } 0 \le \xi \le \xi_\epsilon \\ M_2M_3e^{-\alpha_1(\xi-\xi_\epsilon)} & \text{for } \xi \ge \xi_\epsilon. \end{cases}$$

Proof. The proof is straightforward, using the variation of constants formula. We write the solution of (6C.1) as

$$u(\xi) = \Gamma_1(\xi)\Psi_1(\xi) + \Gamma_2(\xi)\Psi_2(\xi) \tag{6C.4}$$

where Γ_1 and Γ_2 are as in Proposition 6B.1, and

$$\text{(a)} \quad \Psi_1(\xi) = -\frac{1}{W_0}\int_{\xi_\epsilon}^{\xi} e^{-\epsilon\theta(\eta-\xi_\epsilon)}\Gamma_2(\eta)g(\eta)d\eta,$$

$$\text{(b)} \quad \Psi_2(\xi) = -\frac{1}{W_0}\int_{\xi}^{\infty} e^{-\epsilon\theta(\eta-\xi_\epsilon)}\Gamma_1(\eta)g(\eta)d\eta, \tag{6C.5}$$

$$\text{(c)} \quad W_0 = \Gamma_1(\xi_\epsilon)\Gamma_2'(\xi_\epsilon) - \Gamma_1'(\xi_\epsilon)\Gamma_2(\xi_\epsilon).$$

The proof now follows from estimating the various terms in (6C.5) using Proposition 6B.1 and the assumptions on $g(\xi)$. For the proof we shall assume that $|\Gamma_1(\xi)| < 2K$ and $|\Gamma_1'(\xi)| \leq 2K$ for all $\xi > 0$, where K was defined in (2A.1). Actually, all we need is that both $\Gamma_1(\xi)$ and $\Gamma_1'(\xi)$ are bounded for $\xi > \xi_\epsilon$. This will be verified in the proof of Proposition 6C.2.

To begin with, note that from (6B.2c,d), we have that

$$|W_0| > 1. \tag{6C.6}$$

We now consider $\Psi_1(\xi)$. From (6B.2b), (6C.2), (6C.6), and Proposition 2E.1, it follows that for $0 \leq \xi \leq \xi_\epsilon$,

$$\begin{aligned} |\Psi_1(\xi)| &\leq \int_\xi^{\xi_\epsilon} e^{\epsilon\theta\xi_\epsilon} M_1 |\ln\epsilon| M_2 d\eta \\ &\leq 2M_1M_2|\ln\epsilon|(\xi_\epsilon - \xi) \end{aligned}$$

if ϵ is sufficiently small. We are assuming that $|\Gamma_1(\xi)| \leq 2K$ for $0 \leq \xi \leq \xi_\epsilon$. Therefore,

$$|\Gamma_1(\xi)\Psi_1(\xi)| \leq 4M_1M_2K|\ln\epsilon|(\xi_\epsilon - \xi) \quad \text{for } 0 \leq \xi \leq \xi_\epsilon \tag{6C.7}$$

and ϵ sufficiently small.

Now suppose that $\xi \geq \xi_\epsilon$. We use (6B.2a), (6C.2), and (6C.6) to conclude that

$$\begin{aligned} |\Psi_1(\xi)| &\leq \left| \int_{\xi_\epsilon}^{\xi} e^{-\epsilon\theta(\eta-\xi_\epsilon)} M_1 e^{\delta_2(\eta-\xi_\epsilon)} M_2 e^{\alpha_1(\eta-\xi_\epsilon)} d\eta \right| \\ &\leq \frac{M_1M_2}{|\delta_2 - \alpha_1 - \epsilon\theta|} |e^{(\delta_2-\alpha_1-\epsilon\theta)(\xi-\xi_\epsilon)}| . \end{aligned}$$

Since $\delta_1 + \delta_2 = \epsilon\theta$ and $\alpha_1 \leq \frac{1}{2}\mathrm{Re}\,(\delta_2 - \epsilon\theta)$, it follows that there exists a constant M_4, which does not depend on ϵ or α_1, such that for ϵ sufficiently small,

$$|\Gamma_1(\xi)\Psi_1(\xi)| \leq M_4M_2e^{-\alpha_1(\xi-\xi_\epsilon)} \quad \text{for } \xi \geq \xi_\epsilon . \tag{6C.8}$$

We now consider $\Psi_2(\xi)$. If $\xi \geq \xi_\epsilon$, then from (6B.2a), (6C.2), (6C.5b), and (6C.6),

$$\begin{aligned} |\Psi_2(\xi)| &\leq \left| \int_\xi^{\infty} e^{-\epsilon\theta(\eta-\xi_\epsilon)} M_1 e^{\delta_1(\eta-\xi_\epsilon)} M_2 e^{-\alpha_1(\eta-\xi_\epsilon)} d\eta \right| \\ &= \frac{M_1M_2}{|\delta_1 - \alpha_1 - \epsilon\theta|} |e^{(\delta_1-\alpha_1-\epsilon\theta)(\xi-\xi_\epsilon)}| . \end{aligned}$$

Since $\delta_1 + \delta_2 = \epsilon\theta$ and $\delta_1 < 0$, it follows from (6B.2a) that M_4 can be chosen so that

$$|\Gamma_2(\xi)\Psi_2(\xi)| \leq M_4M_2e^{-\alpha_1(\xi-\xi_\epsilon)} \quad \text{for } \xi \geq \xi_\epsilon . \tag{6C.9}$$

Finally, consider $\Psi_2(\xi)$ for $0 \leq \xi \leq \xi_\epsilon$. From (6C.2), (6C.6), and our assumption that $|\Gamma_1(\xi)| \leq 2K$, we conclude that

$$\begin{aligned} |\Psi_2(\xi)| &\leq \int_\xi^{\xi_\epsilon} e^{-\epsilon\theta(\eta-\xi_\epsilon)} |\Gamma_1(\eta)g(\eta)| d\eta + |\Psi_2(\xi_\epsilon)| \\ &\leq 2KM_2e^{\epsilon\theta\xi_\epsilon}(\xi_\epsilon - \xi) + M_4M_2 \\ &\leq M_4M_2(\xi_\epsilon - \xi + 1) \end{aligned}$$

if ϵ is sufficiently small. From (6B.2b) we have that

$$|\Gamma_2(\xi)\Psi_2(\xi)| \leq M_1M_2M_4|\ln\epsilon|(\xi_\epsilon - \xi + 1). \tag{6C.10}$$

It now follows from (6C.4), (6C.7)–(6C.10) that M_3 can be chosen independently of ϵ such that for ϵ sufficiently small,

$$|u(\xi)| \leq \begin{cases} M_2M_3|\ln\epsilon|(\xi_\epsilon - \xi + 1) & \text{for } 0 \leq \xi \leq \xi_\epsilon \\ M_2M_3e^{-\alpha_1(\xi-\xi_\epsilon)} & \text{for } \xi \geq \xi_\epsilon. \end{cases} \tag{6C.11}$$

To complete the proof of Proposition 6C.1, we consider $u'(\xi)$. A consequence of the variation of constants formula is that

$$u'(\xi) = \Gamma_1'(\xi)\Psi_1(\xi) + \Gamma_2'(\xi)\Psi_2(\xi).$$

The same assumptions we used for $\Gamma_1(\xi)$ and $\Gamma_2(\xi)$ to estimate $u(\xi)$ also hold for $\Gamma_1'(\xi)$ and $\Gamma_2'(\xi)$. Therefore, (6C.11) holds with $u(\xi)$ replaced with $u'(\xi)$, and the proof of Proposition 6C.1 is complete.

We conclude this section with a result which will be needed when we consider e_2^+. For this result we let $p(\xi)$ be as before, and $\Gamma_1(\xi)$ be as in Proposition 6B.1.

PROPOSITION 6C.2. *There exist positive constants M_0 and α_0, which do not depend on ϵ, such that for ϵ sufficiently small,*

$$\max\{|\Gamma_1(\xi) - p'(\xi)|, \quad |\Gamma_1'(\xi) - p''(\xi)|\} \leq \begin{cases} M_0\epsilon^2|\ln\epsilon|^3(\xi_\epsilon - \xi + 1) & \text{for } 0 \leq \xi \leq \xi_\epsilon \\ M_0\epsilon^2e^{-\alpha_0(\xi-\xi_\epsilon)} & \text{for } \xi \geq \xi_\epsilon. \end{cases}$$

Proof. Let $u(\xi) = \Gamma_1(\xi) - p'(\xi)$, and $\Phi(\xi) = p(\xi) + \Lambda^{-1}q(\xi)$. Differentiating the first equation in (2B.2) we find that

$$u'' - \epsilon\theta u' + (q - \Lambda)\varphi(p)u = \Lambda\Phi'(\xi)\phi(p) - \epsilon^2\lambda\Gamma_1 \equiv g(\xi). \tag{6C.12}$$

Now $\varphi(p)$ is bounded, and from Propositions 2B.2 and 2E.2, there exists a constant k_1, which does not depend on ϵ, such that

$$|\Phi'(\xi)| \leq \begin{cases} \epsilon^2k_1(\ln\epsilon)^2 & \text{for } 0 \leq \xi \leq \xi_\epsilon \\ \epsilon^2k_1e^{-\alpha(\xi-\xi_\epsilon)} & \text{for } \xi \geq \xi_\epsilon. \end{cases}$$

Therefore, we may choose k_0 so that

$$|g(\xi)| \leq \begin{cases} \epsilon^2k_0(\ln\epsilon)^2 & \text{for } 0 \leq \xi \leq \xi_\epsilon \\ \epsilon^2k_0e^{-\alpha(\xi-\xi_\epsilon)} & \text{for } \xi \geq \xi_\epsilon. \end{cases}$$

Now Proposition 6C.2 follows from Proposition 6C.1 as long as $|\Gamma_1(\xi)| < 2K$ and $|\Gamma_1'(\xi)| \leq 2K$. Recall that we assumed this to be true for the proof of Proposition 6C.1. However, it follows from (2A.1) that $|p(\xi)| < K$ and $|p'(\xi)| = |T_z(z)| < K$ for all ξ. The completion of the proof now follows easily.

D. Proof of Proposition 4B.1. Throughout this section we use the notation of §4B. Note that $(u_2(\xi), v_2(\xi))$ satisfies the system:

$$\begin{aligned} &u_2'' - \epsilon\theta u_2' + q\phi(p)u_2 + \phi(p)v_2 = (\sigma + k^2)(p' + \epsilon^2u_2), \\ &\Lambda^{-1}v_2'' - \epsilon\theta v_2' - q\phi(p)u_2 - \phi(p)v_2 = (\sigma + \Lambda^{-1}k^2)(q' + \epsilon^2v_2) \end{aligned} \tag{6.D1}$$

where differentiation is with respect to ξ. Using a straightforward iteration argument, we prove the existence of a solution of (6D.1) which satisfies (4B.4). The estimates needed for the existence proof will prove Proposition 4B.1.

To set up the iteration scheme, we first write (6D.1) as

(6D.2)
$$\begin{aligned} u_2'' - \epsilon\theta u_2' + q\phi(p)u_2 + \phi(p)v_2 &= (\sigma + k^2)p' + \epsilon^2(\sigma + k^2)u_2, \\ \Lambda^{-1}v_2'' - \epsilon\theta\Lambda^{-1}v_2' - q\phi(p)u_2 - \phi(p)v_2 &= (\sigma + k^2)\Lambda^{-1}q' + \epsilon^2(\sigma + \Lambda^{-1}k^2)v_2 \\ &\quad + \epsilon\theta(1 - \Lambda^{-1})v_2' + \sigma(1 - \Lambda^{-1})q' \ . \end{aligned}$$

We set $u^0(\xi) \equiv 0$ and $v^0(\xi) \equiv 0$. Assuming that $(u^{j-1}(\xi), v^{j-1}(\xi))$ have been defined, we let $(u^j(\xi), v^j(\xi))$ be the solution of

(6D.3)
$$\begin{aligned} u^j_{\xi\xi} - \epsilon\theta u^j_\epsilon + q\phi(p)u^j + \phi(p)v^j &= (\sigma + k^2)p_\xi + \epsilon^2(\sigma + k^2)u^{j-1} \\ \Lambda^{-1}v^j_{\xi\xi} - \epsilon\theta\Lambda^{-1}v^j_\xi - q\phi(p)u^j - \phi(p)v^j &= (\sigma + k^2)\Lambda^{-1}q_\xi + \epsilon^2(\sigma + \Lambda^{-1}k^2)v^{j-1} \\ &\quad + \epsilon\theta(1 - \Lambda^{-1})v^{j-1}_\xi + \sigma(1 - \Lambda^{-1})q_\xi \ , \end{aligned}$$

together with the boundary conditions $\lim_{\xi\to\infty}(u^j(\xi), v^j(\xi)) = (0, 0)$.

Let $w^j(\xi) = u^j(\xi) + \Lambda^{-1}v^j(\xi)$. Adding the equations in (6D.3), we find that

(6D.4)
$$\begin{aligned} w^j_{\xi\xi} - \epsilon\theta w^j_\xi &= (\sigma + k^2)(p_\xi + \Lambda^{-1}q_\xi) + \epsilon^2(\sigma + \Lambda^{-1}k^2)u^{j-1} \\ &\quad + \epsilon^2(\sigma + \Lambda^{-1}k^2)v^{j-1} + \epsilon\theta(1 - \Lambda^{-1})v^{j-1}_\xi + \sigma(1 - \Lambda^{-1})q_\xi \\ &\equiv h^{j-1}(\xi) \, . \end{aligned}$$

Let

$$H(\xi) = \begin{cases} c|\ln\epsilon|(\xi_\epsilon - \xi + 1) & \text{for } 0 \le \xi \le \xi_\epsilon \\ ce^{-r(\xi - \xi_\epsilon)} & \text{for } \xi_\epsilon < \xi \end{cases}$$

where the constants c and r are to be determined. Let

$$\mathcal{S} = \{\phi(\xi) : \max(\phi(\xi), \phi'(\xi)) \le \mathrm{H}(\xi) \quad \text{for } \xi \ge 0\} \, .$$

Using induction, we shall prove that $\{u^j, v^j\} \subset \mathcal{S}$ for each j. Clearly, $\{u^0, v^0\} \subset \mathcal{S}$, so assume that $\{u^{j-1}, v^{j-1}\} \subset \mathcal{S}$. We first estimate $h^{j-1}(\xi)$, which was defined in (6D.4), and then $w^j(\xi)$.

In order to estimate $h^{j-1}(\xi)$ we note that $q_\xi(\xi) = Y_z(z)$. From (2A.1b) it follows that $|q_\xi(\xi)| \le K$ for all ξ. This, together with (1.15) and Propositions 2B.2 and 2E.2 imply that there exists a constant c_1 such that if $r \le \alpha$, which we always assume to be true, then

(6D.5)
$$|h^{j-1}(\xi)| \le \begin{cases} \epsilon c_1 & \text{for } 0 \le \xi \le \xi_\epsilon \\ \epsilon c_1 e^{-r(\xi - \xi_\epsilon)} & \text{for } \xi \ge \xi_\epsilon. \end{cases}$$

Throughout this section the constants c_j, $j = 1, 2, \cdots$, do not depend on ϵ.

We now estimate $w^j(\xi)$. From (6D.4), we have that

(6D.6)
$$w^j_\xi(\xi) = e^{\epsilon\theta(\xi - \xi_\epsilon)} \int_\xi^\infty e^{-\epsilon\theta(\eta - \xi_\epsilon)} h^{j-1}(\eta) d\eta.$$

It follows from (6D.5) and (6D.6) that there exists a constant c_2 such that

$$|w_\xi^j(\xi)| \le \begin{cases} \epsilon c_2(\xi_\epsilon - \xi + 1) & \text{for } 0 \le \xi \le \xi_\epsilon \\ \epsilon c_2 e^{-r(\xi-\xi_\epsilon)} & \text{for } \xi \ge \xi_\epsilon. \end{cases} \tag{6D.7}$$

Since $w^j(\xi) = -\int_\xi^\infty w_\xi^j(\eta)d\eta$, we conclude from Proposition 2E.1 that there exists a constant c_3 such that

$$|w^j(\xi)| \le \begin{cases} \epsilon c_3|\ln\epsilon|(\xi_\epsilon - \xi + 1) & \text{for } 0 \le \xi \le \xi_\epsilon \\ \epsilon c_3 e^{-r(\xi-\xi_\epsilon)} & \text{for } \xi \ge \xi_\epsilon. \end{cases} \tag{6D.8}$$

Note that c_3 does not depend on ϵ. It also does not depend on the constant c used in the definition of $\mathcal{S}$ if we assume that $\epsilon c < 1$.

We now consider $(u^j(\xi), v^j(\xi))$. Note that $v^j = -\Lambda u^j + \Lambda w^j$. We plug this into the first equation of (6D.3) to obtain

$$\begin{aligned} u_{\xi\xi}^j - \epsilon\theta u_\xi^j + (q-\Lambda)\phi(p)u^j &= -\Lambda\phi(p)w^j + (\sigma + k^2)p_\xi + \epsilon^2(\sigma+k^2)u^{j-1} \\ &\equiv g^j(\xi)\,. \end{aligned} \tag{6D.9}$$

We wish to apply Proposition 6C.1, with $\lambda = 0$, to obtain the desired solution of (6D.9). To do this we must first estimate $g^j(\xi)$. From (6D.8), Proposition 2E.1, (1.15), (2A.1), and the fact that $u^{j-1} \in \mathcal{S}$, we conclude that M_2 can be chosen so that $g_j(z)$ satisfies (6C.2) with $\alpha_1 = r$. Note that M_2 does not depend on the constant c in the definition of $\mathcal{S}$, if we assume, as before, that $\epsilon c < 1$. We also assume that $r \le \min\{\alpha, \frac{1}{2}\mathrm{Re}\,(\delta_2 - \epsilon\theta)\}$. We may then apply Proposition 6C.1 to conclude that there exists a constant M_4, which does not depend on ϵ, such that for ϵ sufficiently small, there exists a solution $u^j(\xi)$ of (6D.9) with $\lim_{\xi\to\infty} u^j(\xi) = 0$, and

$$\max\{|u^j(\xi)|, |u_\xi^j(\xi)|\} \le \begin{cases} M_4|\ln\epsilon|(\xi_\epsilon - \xi + 1) & \text{for } 0 \le \xi \le \xi_\epsilon \\ M_4 e^{-r(\xi-\xi_\epsilon)} & \text{for } \xi \ge \xi_\epsilon. \end{cases} \tag{6D.10}$$

Hence, $u^j(\xi) \in \mathcal{S}$ if $M_4 \le c$. Because $v^j = -\Lambda u^j + \Lambda w^j$, we conclude from (6D.10) and (6D.8) that c can be chosen so that $v^j(\xi) \in \mathcal{S}$.

We now complete the proof of Proposition 4B.1 by showing that $\{(u^j, v^j)\}$ forms a Cauchy sequence. It is then straightforward to show that some subsequence of $\{(u^j, v^j)\}$ converges to a solution of (6D.1) which satisfies (4B.9).

Let $H(\xi)$ be as in the definition of $\mathcal{S}$. We shall prove, by induction, that

$$|u^j(\xi) - u^{j-1}(\xi)| + |v^j(\xi) - v^{j-1}(\xi)| \le \epsilon^{j-1}H(\xi) \tag{6D.11}$$

for $\xi \ge 0$ and $j = 1, 2, \cdots$. Since $u^0(\xi) \equiv 0 \equiv v^0(\xi)$, and $\{u^1(\xi), v^1(\xi)\} \subset \mathcal{S}$, (6D.11) holds for $j = 1$. We assume that it holds for some $j \ge 1$, and prove it holds for $j+1$.

From (6C.5) we have that for each $j \ge 1$, $u^j = \Gamma_1\Psi_1^j + \Gamma_2\Psi_2^j$, where

$$\begin{aligned} \Psi_1^j(\xi) &= -\frac{1}{W_0}\int_{\xi_\epsilon}^{\xi} e^{-\epsilon\theta(\eta-\xi_\epsilon)}\Gamma_2(\eta)g^j(\eta)d\eta\,, \\ \Psi_2^j(\xi) &= -\frac{1}{W_0}\int_{\xi}^{\infty} e^{-\epsilon\theta(\eta-\xi_\epsilon)}\Gamma_1(\eta)g^j(\eta)d\eta\,, \end{aligned} \tag{6D.12}$$

$\Gamma_1(\xi)$ and $\Gamma_2(\xi)$ are as in Proposition 6B.1 with $\lambda = 0$, and $g^j(\xi)$ is defined in (6D.9). We must now do quite a bit of estimating. Recalling (6D.4) and (6D.6), we begin with

$$w_\xi^{j+1}(\xi) - w_\xi^j(\xi) = e^{\epsilon\theta(\xi-\xi_\epsilon)} \int_\xi^\infty e^{-\epsilon\theta(\eta-\xi_\epsilon)} [h^j(\eta) - h^{j-1}(\eta)]\, d\eta \tag{6D.13}$$

where

$$\begin{aligned} h^j(\eta) - h^{j-1}(\eta) &= \epsilon^2(\sigma + \Lambda^{-1}k^2)(u^j - u^{j-1}) + \epsilon^2(\sigma + \Lambda^{-1}k^2)(v^j - v^{j-1}) \\ &\quad + \epsilon\theta(1 - \Lambda^{-1})(v_\xi^j - v_\xi^{j-1})\,. \end{aligned} \tag{6D.14}$$

We plug (6D.14) into (6D.13) and integrate by parts to obtain

$$\begin{aligned} &w_\xi^{j+1}(\xi) - w_\xi^j(\xi) \\ &\qquad = e^{\epsilon\theta(\xi-\xi_\epsilon)} \int_\xi^\infty e^{-\epsilon\theta(\eta-\xi_\epsilon)} \{\epsilon^2(\sigma + \Lambda^{-1}k^2)(u^j - u^{j-1} + v^j - v^{j-1})\} d\eta \\ &\qquad\quad - \epsilon\theta(1 - \Lambda^{-1})(v^j(\xi) - v^{j-1}(\xi)) \\ &\qquad\quad + \epsilon^2\theta^2(1 - \Lambda^{-1}) e^{\epsilon\theta(\xi-\xi_\epsilon)} \int_\xi^\infty e^{-\epsilon\theta(\eta-\xi_\epsilon)} (v^j - v^{j-1}) d\eta\,. \end{aligned}$$

Using (1.15), we conclude that there exists a constant c_1 such that for $\xi \geq 0$

$$\begin{aligned} |w_\xi^{j+1}(\xi) - w_\xi^j(\xi)| &\leq \epsilon^2 c_1 |v^j(\xi) - v^{j-1}(\xi)| \\ &\quad + \epsilon^2 c_1 \int_\xi^\infty |u^j(\eta) - u^{j-1}(\eta)| + |v^j(\eta) - v^{j-1}(\eta)| d\eta\,. \end{aligned}$$

Therefore,

$$\begin{aligned} |w^{j+1}(\xi) - w^j(\xi)| &\leq \int_\xi^\infty |w_\xi^{j+1}(\eta) - w_\xi^j(\eta)| d\eta \\ &\leq \epsilon^2 c_1 \int_\xi^\infty |v^j(\eta) - v^{j-1}(\eta)| d\eta \\ &\quad + \epsilon^2 c_1 \int_\xi^\infty \int_\eta^\infty |u^j(s) - u^{j-1}(s)| + |v^j(s) - v^{j-1}(s)| ds\, d\eta \\ &\leq \epsilon^2 c_1 \int_\xi^\infty |v^j(\eta) - v^{j-1}(\eta)| d\eta \\ &\quad + \epsilon^2 c_1 \int_\xi^\infty (\eta - \xi)(|u^j(\eta) - u^{j-1}(\eta)| + |v^j(\eta) - v^{j-1}(\eta)|) d\eta\,. \end{aligned}$$

This, together with the definition of $g^j(\xi)$ given in (6D.9), and (1.15) imply that there exists a constant c_2 such that for $\xi \geq 0$,

$$\begin{aligned} |g^{j+1}(\xi) - g^j(\xi)| &\leq \epsilon^2 c_2 |u^j(\xi) - u^{j-1}(\xi)| + \epsilon^2 c_2 \int_\xi^\infty |v^j(\eta) - v^{j_1}(\eta)| d\eta \\ &\quad + \epsilon^2 c_2 \int_\xi^\infty (\eta - \xi)(|u^j(\eta) - u^{j-1}(\eta)| + |v^j(\eta) - v^{j-1}(\eta)|) d\eta \\ &\leq \epsilon^{j+1} c_2 H(\xi) + \epsilon^{j+1} c_2 \int_\xi^\infty H(\eta)(1 + \eta - \xi) d\eta\,. \end{aligned} \tag{6D.15}$$

A straightforward computation shows that for ϵ sufficiently small and $\xi \geq 0$,

$$\epsilon^{1/2} \int_{\xi}^{\infty} H(\eta)(1+\eta-\xi)d\eta \leq \frac{1}{2} c_2 H(\xi)\,.$$

Together with (6D.15), this implies that if ϵ is sufficiently small and $\xi \geq 0$, then

$$|g^{j+1}(\xi) - g^j(\xi)| \leq \epsilon^{j+1/2} H(\xi)\,. \tag{6D.16}$$

The remainder of the proof is now straightforward. From (6D.12), (6D.16), and the properties of $\Gamma_1(\xi)$ and $\Gamma_2(\xi)$ given in Proposition 6B.1, we find that if ϵ is sufficiently small and $\xi \geq 0$,

$$|\Gamma_1(\xi)(\Psi_1^{j+1}(\xi) - \Psi_1^j(\xi))| \leq \tfrac{1}{4}\epsilon^j H(\xi), \qquad |\Gamma_2(\xi)(\Psi_2^{j+1}(\xi) - \Psi_1^j(\xi))| \leq \tfrac{1}{4}\epsilon^j H(\xi)\,.$$

Therefore,

$$\begin{aligned} |u^{j+1}(\xi) - u^j(\xi)| &\leq |\Gamma_1(\xi)(\Psi_1^{j+1}(\xi) - \Psi_1^j(\xi))| + |\Gamma_2(\xi)(\Psi_2^{j+1}(\xi) - \Psi_2^j(\xi))| \\ &\leq \tfrac{1}{2}\epsilon^j H(\xi)\,. \end{aligned}$$

A similar calculation shows that $|v^{j+1}(\xi) - v^j(\xi)| \leq \frac{1}{2}\epsilon^j H(\xi)$. This completes the proof that $\{(u^j(\xi), v^j(\xi))\}$ forms a Cauchy sequence. As we mentioned earlier, it follows quite easily that some subsequence of $\{(u^j(\xi), v^j(\xi))\}$ converges to a solution of (6D.1) which satisfies (4B.9).

E. Proof of Proposition 4C.1. Throughout this section we use the notation of §4C. The proof of Proposition 4C.1 is similar to the proof of Proposition 4B.1. We prove the existence of a solution of (4C.1) using an iteration scheme. The estimates needed for the iteration scheme will imply that the proposition is true. To set up the iteration scheme we consider the system

$$\begin{aligned} &u^j_{\xi\xi} - \epsilon\theta u^j_{\epsilon} + q\phi(p)u^j + \phi(p)v^j - \epsilon^2(\sigma+k^2)u^j = 0, \\ &\Lambda^{-1}v^j_{\xi\xi} - \epsilon\theta\Lambda^{-1}v^j_{\xi} - q\phi(p)u^j - \phi(p)v^j - \epsilon^2(\sigma+k^2)\Lambda^{-1}v^j \\ &\qquad\qquad = \epsilon\theta(1-\Lambda^{-1})v^{j-1}_{\xi} + \epsilon(1-\Lambda^{-1})\sigma v^{j-1} \end{aligned} \tag{6E.1}$$

for $j \geq 1$. Let $w^j = u^j + \Lambda^{-1}v^j$. Then $w^j(\xi)$ satisfies the equation

$$w^j_{\xi\xi} - \epsilon\theta w^j_{\xi} - \epsilon^2(\sigma+k^2)w^j = \epsilon\theta(1-\Lambda^{-1})v^{j-1}_{\xi} + \epsilon^2(1-\Lambda^{-1})\sigma v^{j-1}\,. \tag{6E.2}$$

We assume that

$$w^j(\xi) = \epsilon e^{\epsilon\gamma\xi} + \epsilon^2\Psi^j(\xi) \tag{6E.3}$$

where γ was defined in (4C.4). Then $\Psi^j(\xi)$ is a solution of the equation

$$\begin{aligned} \Psi^j_{\xi\xi} - \epsilon\theta\Psi^j_{\xi} - \epsilon^2(\sigma+k^2)\Psi^j &= \frac{\theta}{\epsilon}(1-\Lambda^{-1})v^{j-1}_{\xi} + (1-\Lambda^{-1})\sigma v^{j-1} \\ &\equiv h^{j-1}(\xi)\,. \end{aligned} \tag{6E.4}$$

We assume that $u^0(\xi) = v^0(\xi) = 0$ for all ξ.

We now set

$$H(\xi) = \begin{cases} \epsilon c(\ln \epsilon)^2(\xi_\epsilon - \xi + 1) & \text{for } 0 \le \xi \le \xi_\epsilon \\ \epsilon c e^{-\epsilon r(\xi - \xi_\epsilon)} & \text{for } \xi_\epsilon < \xi \end{cases} \tag{6E.5}$$

where r and c are positive constants, which do not depend on ϵ, and will be determined later. Let

$$\mathcal{S} = \{\phi(\xi) : \max(\phi(\xi), \phi'(\xi)) \le \mathrm{H}(\xi) \quad \text{for } \xi \ge 0\}. \tag{6E.6}$$

Using induction we shall prove that $u^j \in \mathcal{S}$ and $v^j \in \mathcal{S}$ for all j. This is certainly true for $j = 0$, so we assume it is true for $j - 1$. We must now estimate $w^j(\xi)$. This is, unfortunately, a rather long and involved calculation.

Using the variation of constants formula we write $\Psi^j(\xi)$ as

$$\Psi^j(\xi) = e^{\epsilon\gamma(\xi - \xi_\epsilon)}\Psi_1(\xi) + e^{\epsilon\delta(\xi - \xi_\epsilon)}\Psi_2(\xi) \tag{6E.7}$$

where γ is as before, and

$$\text{(a)} \quad \delta = \frac{\theta + \sqrt{\theta^2 + 4(\sigma + k^2)}}{2},$$

$$\text{(b)} \quad \Psi_1(\xi) = -\frac{1}{\epsilon\sqrt{\theta^2 + 4(\sigma + k^2)}} \int_{\xi_\epsilon}^{\xi} e^{\epsilon(\delta - \theta)(\eta - \xi_\epsilon)} h^{j-1}(\eta) d\eta, \tag{6E.8}$$

$$\text{(c)} \quad \Psi_2(\xi) = -\frac{1}{\epsilon\sqrt{\theta^2 + 4(\sigma + k^2)}} \int_{\xi}^{\infty} e^{\epsilon(\gamma - \theta)(\eta - \xi_\epsilon)} h^{j-1}(\eta) d\eta.$$

First we estimate $\Psi_1(\xi)$. Using the definition of h^{j-1} given in (6E.4) we have that

$$\begin{aligned} \Psi_1(\xi) = &- \frac{\theta(1 - \Lambda^{-1})}{\epsilon^2\sqrt{\theta^2 + 4(\sigma + k^2)}} \int_{\xi_\epsilon}^{\xi} e^{\epsilon(\delta - \theta)(\eta - \xi_\epsilon)} v_\xi^{j-1}(\eta) d\eta \\ &- \frac{\theta(1 - \Lambda^{-1})\sigma}{\epsilon\sqrt{\theta^2 + 4(\sigma + k^2)}} \int_{\xi_\epsilon}^{\xi} e^{\epsilon(\delta - \theta)(\eta - \xi_\epsilon)} v^{j-1}(\eta) d\eta \\ &= \text{(I)} + \text{(II)}. \end{aligned} \tag{6E.9}$$

We integrate by parts to obtain (I) = (Ia) + (Ib) where

$$\text{(Ia)} = \frac{-\theta(1 - \Lambda^{-1})}{\epsilon^2\sqrt{\theta^2 + 4(\sigma + \mathrm{k}^2)}} \left[\mathrm{e}^{\epsilon(\delta - \theta)(\xi - \xi_\epsilon)}\mathrm{v}^{\mathrm{j}-1}(\xi) - \mathrm{v}^{\mathrm{j}-1}(\xi_\epsilon)\right]$$

and

$$\text{(Ib)} = \frac{-\theta(1 - \Lambda^{-1})(\theta - \delta)}{\epsilon\sqrt{\theta^2 + 4(\sigma + \mathrm{k}^2)}} \int_{\xi_\epsilon}^{\xi} \mathrm{e}^{\epsilon(\delta - \theta)(\eta - \xi_\epsilon)}\mathrm{v}^{\mathrm{j}-1}(\eta)\mathrm{d}\eta.$$

Suppose that $0 \le \xi \le \xi_\epsilon$. It follows from (1.15) and the assumption that $v^{j-1} \in \mathcal{S}$ that

$$\begin{aligned} |\text{(Ia)}| &\le \frac{\theta M}{\epsilon\theta}(\epsilon c(\ln \epsilon)^2(\xi_\epsilon - \xi + 1) + \epsilon c) \\ &\le c c_1(\ln \epsilon)^2(\xi_\epsilon - \xi + 1) \end{aligned} \tag{6E.10}$$

for some constant c_1 which does not depend on ϵ. It follows from (1.15) and Proposition 2E.1 that if $0 \le \xi \le \xi_\epsilon$, then

$$\begin{aligned} |(\mathrm{Ib})| &\le \frac{\theta M\delta}{\theta}\int_{\xi_\epsilon}^{\xi} \epsilon c(\ln\epsilon)^2(\xi_\epsilon - \eta + 1)\mathrm{d}\eta \\ &\le cc_1(\ln\epsilon)^2(\xi_\epsilon - \xi + 1) \end{aligned} \tag{6E.11}$$

if ϵ is sufficiently small.

Now suppose that $\xi \ge \xi_\epsilon$. Using (1.15) and the assumption that $v^{j-1} \in \mathcal{S}$, we have that

$$\begin{aligned} |(\mathrm{Ia})| &\le \frac{\theta M}{\epsilon\theta}\left|e^{\epsilon(\delta-\theta)(\xi-\xi_\epsilon)}\epsilon c e^{-\epsilon r(\xi-\xi_\epsilon)} + \epsilon c\right| \\ &\le 2Mc\left|e^{\epsilon(\delta-\theta-r)(\xi-\xi_\epsilon)}\right| \end{aligned} \tag{6E.11}$$

where we assume that

$$r \le \mathrm{Re}\,\frac{\delta-\theta}{2} = \mathrm{Re}\,\frac{\sqrt{\theta^2 + 4(\sigma + \mathrm{k}^2)} - \theta}{2}. \tag{6E.12}$$

From (1.14), we may choose $r > 0$ independent of ϵ. On the other hand,

$$\begin{aligned} |(\mathrm{Ib})| &\le \frac{\theta M\delta}{\theta}\left|\int_{\xi_\epsilon}^{\xi} e^{\epsilon(\delta-\theta)(\eta-\xi_\epsilon)}\epsilon c e^{-\epsilon r(\eta-\xi_\epsilon)}d\eta\right| \\ &\le \frac{M\delta c}{|\delta-\theta-r|}\left|e^{\epsilon(\delta-\theta-r)(\xi-\xi_\epsilon)}\right|. \end{aligned} \tag{6E.13}$$

Combining (6E.9)–(6E.13) we conclude that there exists a constant c_2, which does not depend on ϵ, such that

$$|(\mathrm{I})| \le \begin{cases} cc_2(\ln\epsilon)^2(\xi_\epsilon - \xi + 1) & \text{for } 0 \le \xi \le \xi_\epsilon \\ cc_2\left|e^{\epsilon(\delta-\theta-r)(\xi-\xi_\epsilon)}\right| & \text{for } \xi \ge \xi_\epsilon. \end{cases} \tag{6E.14}$$

The estimates for (II) are easier. If $0 \le \xi \le \xi_\epsilon$, then from (1.15), Proposition 2E.1, and the assumption that $v^{j-1} \in \mathcal{S}$ we conclude that

$$\begin{aligned} |(\mathrm{II})| &\le \frac{M^2}{\theta}\int_{\xi_\epsilon}^{\xi} \epsilon c(\ln\epsilon)^2(\xi_\epsilon - \eta + 1)\mathrm{d}\eta \\ &\le 1 \end{aligned} \tag{6E.15}$$

if ϵ is sufficiently small. If $\xi \ge \xi_\epsilon$, then using (1.15) and the assumption that $v^{j-1} \in \mathcal{S}$, we find that

$$\begin{aligned} |(\mathrm{II})| &\le \frac{M^2}{\theta}\left|\int_{\xi_\epsilon}^{\xi} e^{\epsilon(\delta-\theta)(\eta-\xi_\epsilon)}\epsilon c e^{-r(\eta-\xi_\epsilon)}d\eta\right| \\ &\le \left|\frac{M^2 c}{\theta(\delta-\theta-r)}e^{\epsilon(\delta-\theta-r)(\xi-\xi_\epsilon)}\right|. \end{aligned} \tag{6E.16}$$

Combining (6E.14)–(6E.16), we conclude that there exists a constant c_3, which does not depend on ϵ, such that

$$|\Psi_1(\xi)| \le \begin{cases} \frac{1}{2}c_3 c(\ln\epsilon)^2(\xi_\epsilon - \xi + 1) & \text{for } 0 \le \xi \le \xi_\epsilon \\ \frac{1}{2}c_3 c|e^{\epsilon(\delta-\theta-r)(\xi_\epsilon-\xi)}| & \text{for } \xi \ge \xi_\epsilon. \end{cases}$$

Since $\delta + \gamma = \theta$, it follows that if ϵ is sufficiently small, then

$$|e^{\epsilon\gamma(\xi-\xi_\epsilon)}\Psi_1(\xi)| \le \begin{cases} \frac{1}{2}c_3 c(\ln\epsilon)^2(\xi_\epsilon - \xi + 1) & \text{for } 0 \le \xi \le \xi_\epsilon \\ \frac{1}{2}c_3 c e^{-\epsilon r(\xi-\xi_\epsilon)} & \text{for } \xi \ge \xi_\epsilon\,. \end{cases} \tag{6E.17}$$

A similar estimate holds for $\Psi_2(\xi)$. That is, c_3 can be chosen so that

$$|e^{\epsilon\delta(\xi-\xi_\epsilon)}\Psi_2(\xi)| \le \begin{cases} \frac{1}{2}c_3 c(\ln\epsilon)^2(\xi_\epsilon - \xi + 1) & \text{for } 0 \le \xi \le \xi_\epsilon \\ \frac{1}{2}c_3 c e^{-\epsilon r(\xi-\xi_\epsilon)} & \text{for } \xi \ge \xi_\epsilon \end{cases}$$

if ϵ is sufficiently small. From (6E.7), we conclude that if ϵ is sufficiently small, then

$$|\Psi^j(\xi)| \le \begin{cases} c_3 c(\ln\epsilon)^2(\xi_\epsilon - \xi + 1) & \text{for } 0 \le \xi \le \xi_\epsilon \\ c_3 c e^{-\epsilon r(\xi-\xi_\epsilon)} & \text{for } \xi \ge \xi_\epsilon. \end{cases} \tag{6E.18}$$

Note that $\Psi^j_\xi(\xi) = \epsilon\gamma e^{\epsilon\gamma(\xi-\xi_\epsilon)}\Psi_1(\xi) + \epsilon\delta e^{\epsilon\delta(\xi-\xi_\epsilon)}\Psi_2(\xi)$. From our estimates on $\Psi_1(\xi)$ and $\Psi_2(\xi)$ it follows that c_3 can be chosen so that

$$|\Psi^j_\xi(\xi)| \le \begin{cases} \epsilon c_3 c(\ln\epsilon)^2(\xi_\epsilon - \xi + 1) & \text{for } 0 \le \xi \le \xi_\epsilon \\ \epsilon c_3 c e^{-\epsilon r(\xi-\xi_\epsilon)} & \text{for } \xi \ge \xi_\epsilon. \end{cases} \tag{6E.19}$$

Recall that $v^j = -\Lambda w^j + \Lambda w^j$. We plug this into the first equation of (6E.1) to obtain

$$u^j_{\xi\xi} - \epsilon\theta u^j_\xi + [(q-\Lambda)\phi(p) - \epsilon^2(\sigma + k^2)]\,u^j = -\Lambda\phi(p)w^j.$$

We wish to apply Proposition 6C.1 with $\lambda = -(\sigma + k^2)$, $u(\xi) = u^j(\xi)/\epsilon$, and $g(\xi) = -\Lambda\phi(p)w^j(\xi)/\epsilon$. Then, from (1.15), $\lambda_0 = |\lambda| \le 2M^2$, assuming that $M > 1$. Moreover, it follows from (6E.4), (6E.18), and Proposition 2E.1, that if $0 \le \xi \le \xi_\epsilon$, and ϵ is sufficiently small, then

$$\begin{aligned} |g(\xi)| &\le 2|e^{\epsilon\gamma\xi}| + 2\epsilon|\Psi^j(\xi)| \le 2|e^{\epsilon\gamma\xi_\epsilon}| + 2\epsilon c_3 c(\ln\epsilon)^2(\xi_\epsilon + 1) \\ &\le 5\,. \end{aligned}$$

If $\xi \ge \xi_\epsilon$ and ϵ is sufficiently small, then from (6E.4) and (6E.18),

$$\begin{aligned} |g(\xi)| &\le 2|e^{\epsilon\gamma\xi}| + 2\epsilon c_3 c|e^{-\epsilon r(\xi-\xi_\epsilon)}| \\ &\le 5e^{-\epsilon r(\xi-\xi_\epsilon)}\,. \end{aligned}$$

We now apply Proposition 6C.1 with $M_2 = 5$ and $\alpha_1 = \epsilon r$ to conclude that there exists a constant M_3, which does not depend on ϵ such that

$$\max\{|u^j(\xi)|, |u^j_\xi(\xi)|\} \le \begin{cases} \epsilon M_3|\ln\epsilon|(\xi_\epsilon - \xi + 1) & \text{for } 0 \le \xi \le \xi_\epsilon \\ \epsilon M_3 e^{-r(\xi-\xi_\epsilon)} & \text{for } \xi \ge \xi_\epsilon. \end{cases}$$

Therefore, $w^j(\xi) \in \mathcal{S}$. Since $v^j = -\Lambda u^j + \Lambda w^j$, it easily follows that $v^j \in \mathcal{S}$.

Recall from the statement of Proposition 4C.1 that we must estimate $u(0) = P_2$ and $u_x(0) = P_3$. For this reason, we estimate $w^j(0)$ and $u^j_\xi(0)$. As in (6C.4) and (6C.5), we find that

$$w^j(\xi) = \Gamma_1(\xi)\Psi_1(\xi) + \Gamma_2(\xi)\Psi_2(\xi) \tag{6E.20}$$

where Γ_1 and Γ_2 are as in Proposition 6B.1, and

$$\begin{aligned}
&\text{(a)} \quad \Psi_1(\xi) = -\frac{\Lambda}{W_0}\int_{\xi_\epsilon}^{\xi} e^{-\epsilon\theta(\eta-\xi_\epsilon)}\Gamma_2(\eta)\phi(p)w^j(\eta)d\eta\,,\\
&\text{(b)} \quad \Psi_2(\xi) = -\frac{\Lambda}{W_0}\int_{\xi}^{\infty} e^{-\epsilon\theta(\eta-\xi_\epsilon)}\Gamma_1(\eta)\phi(p)w^j(\eta)d\eta\,,\\
&\text{(c)} \quad W_0 \equiv \Gamma_1(\xi_\epsilon)\Gamma_2'(\xi_\epsilon) - \Gamma_1'(\xi_\epsilon)\Gamma_2(\xi_\epsilon)\,.
\end{aligned} \tag{6E.21}$$

Recall from (6C.6) that $|W_0| > 1$. It then follows from (6B.2b), (6E.3), (6E.18), and Proposition 2E.1 that for ϵ sufficiently small,

$$\begin{aligned}
|\Psi_1(0)| &\le \Lambda\int_0^{\xi_\epsilon} e^{-\epsilon\theta(\eta-\xi_\epsilon)}M_1|\ln\epsilon|2\epsilon e^{\epsilon\gamma\eta}d\eta\\
&\le (2\Lambda M_1)\epsilon|\ln\epsilon|e^{\epsilon(\gamma+\theta)\xi_\epsilon}\xi_\epsilon\\
&\le N_1\epsilon(\ln\epsilon)^2
\end{aligned} \tag{6E.22}$$

for some constant N_1 which does not depend on ϵ. From Proposition 6C.2, there exists a constant N_2 such that

$$|\Gamma_1(\xi)| \le N_2 \quad \text{for } 0 \le \xi \le \xi_\epsilon.$$

It therefore follows that

$$|\Gamma_1(0)\Psi_1(0)| \le N_1N_2\epsilon(\ln\epsilon)^2\,.$$

A similar estimate holds for $|\Gamma_2(0)\Psi_2(0)|$. From (6E.20) we conclude that c_1 can be chosen so that

$$|w^j(0)| \le c_1\epsilon(\ln\epsilon)^2\,. \tag{6E.23}$$

We must now consider

$$u^j_\xi(0) = \Gamma_1'(0)\Psi_1(0) + \Gamma_2'(0)\Psi_2(0)\,. \tag{6E.24}$$

From Propositions 6B.1 and 6B.3, Corollary 2D.2, and (2D.3), it follows that

$$\begin{aligned}
&\text{(a)} \quad \Gamma_1(0) = p'(0) + \text{ h.o.t. } = \theta + 0(\epsilon\ln\epsilon)\\
&\text{(b)} \quad \Gamma_1'(0) = p''(0) + \text{ h.o.t. } = \epsilon\theta^2 + \text{ h.o.t.}\\
&\text{(c)} \quad |\Gamma_2(0)| \le M_1|\ln\epsilon|\\
&\text{(d)} \quad |\Gamma_2'(0)| \ge M^1.
\end{aligned} \tag{6E.25}$$

It now follow from (6E.22) and (6E.25b) that

$$|\Gamma_1'(0)\Psi_1(0)| \le \epsilon^2\theta^2N_1(\ln\epsilon)^2\,. \tag{6E.26}$$

We now consider $\Gamma_2(0)\Psi_2(0)$. We write $\Psi_2(0)$ as

$$\Psi_2(0) = \frac{\Lambda e^{\epsilon\theta\xi_\epsilon}}{W_0}\int_0^\infty e^{-\epsilon\theta\eta}\Gamma_1(\eta)\phi(p)w^j(\eta)d\eta. \tag{6E.27}$$

Note that

$$\begin{aligned}\int_0^\infty & e^{-\epsilon\theta\eta}\Gamma_1(\eta)\phi(p)w^j(\eta)d\eta \\ &= \epsilon\int_0^\infty p'(\eta)\phi(p)d\eta + \epsilon\int_0^\infty (\Gamma_1(\eta) - p'(\eta))\,\phi(p)d\eta \\ &\quad + \int_0^\infty (e^{-\epsilon\theta\eta}w^j(\eta) - \epsilon)\,\Gamma_1(\eta)\phi(p)d\eta \\ &= (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}).\end{aligned}$$

Now,

$$\begin{aligned}(\mathrm{I}) &= \epsilon\int_0^\infty e^p p'(\eta)d\eta = \epsilon\int_{p(0)}^0 e^p dp \\ &= \epsilon - \epsilon e^{p(0)} \\ &= \epsilon - \epsilon^7.\end{aligned}$$

From Proposition 6C.2, it follows that if ϵ is sufficiently small, then $|(\mathrm{II})| \leq \epsilon^2$. It remains to consider (III). A straightforward, but tedious computation shows that if we use the bounds (6E.3) and (6E.18) for $w^j(\eta)$, the bounds given by Proposition 6C.2 for $\Gamma_1(\eta)$, and the fact that $|\phi(p)| < 1$ for all η, then for some constant N_0, independent of ϵ

$$|(\mathrm{III})| \leq N_0\epsilon^2(\ln\epsilon)^4.$$

Therefore, if ϵ is sufficiently small, then

$$\int_0^\infty e^{-\epsilon\theta\eta}\Gamma_1(\eta)\phi(p)w^j(\eta)d\eta = \epsilon + O\left(\epsilon^2(\ln\epsilon)^4\right). \tag{6E.28}$$

We now estimate $e^{\epsilon\theta\xi_\epsilon}/W_0$ which appears in the formula for $\Psi_2(0)$ in (6E.27). It is straightforward to check that if

$$W(\xi) = \Gamma_1(\xi)\Gamma_2'(\xi) - \Gamma_1'(\xi)\Gamma_2(\xi),$$

then $W' = \epsilon\theta W$. Therefore, $W(\xi) = W_0 e^{\epsilon\theta(\xi-\xi_\epsilon)}$. It follows that

$$\frac{e^{\epsilon\theta\xi_\epsilon}}{W_0} = \frac{1}{W(0)}.$$

This, together with (6E.25) implies that

$$\begin{aligned}\frac{\Gamma_2'(0)e^{\epsilon\theta\xi_\epsilon}}{W_0} &= \frac{\Gamma_2'(0)}{\Gamma_1(0)\Gamma_2'(0) - \Gamma_1'(0)\Gamma_2(0)} \\ &= \frac{\Gamma_2'(0)}{\theta\Gamma_2'(0) + O(\epsilon\ln\epsilon)} = \frac{1}{\theta} + O(\epsilon\ln\epsilon).\end{aligned} \tag{6E.29}$$

Combining (6E.27)–(6E.29), we have that

$$\Gamma_2'(0)\Psi_2(0) = \frac{\epsilon}{\theta} + O(\epsilon^2(\ln\epsilon)^4) . \tag{6E.30}$$

It then follows from (6E.24), (6E.26), and (6E.30) that

$$u_\xi^j(0) = \frac{\epsilon}{\theta} + O(\epsilon^2(\ln\epsilon)^4) .$$

This then implies that

$$u_x^j(0) = \frac{1}{\theta} + O(\epsilon(\ln\epsilon)^4) . \tag{6E.31}$$

We have now shown that for each j, $w^j(\xi) \in \mathcal{S}$, $v^j \in \mathcal{S}$, and (6E.23) and (6E.31) are satisfied. To complete the proof of Proposition 4C.1, we must demonstrate that some subsequence of $\{(w^j(\xi), v^j(\xi))\}$ converges to a solution of (4C.1). The proof of this last statement is similar to the proof given in §6D. Since that argument was worked out in complete detail, we do not give any details here.

REFERENCES

[1] H. Berestycki, B. Nicolaenko, and B. Scheurer, *Traveling wave solutions to combustion models and their singular limits*, SIAM J. Math. Anal., 6 (1985), pp. 1207–1242.

[2] E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[3] J. W. Evans, *Nerve axon equations* IV: *The stable and unstable impulse*, Indiana University Math. J., 24 (1975), pp. 1169–1190.

[4] P. C. Fife, *Nonlinear Diffusive Waves*, preprint.

[5] R. Gardner and C. Jones, *Private communication.*

[6] D. Henry, *The Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, New York, 1981.

[7] G. Joulin and P. Clavin, *Linear stability analysis of nonadiabatic flames: diffusional-thermal model*, Combust. and Flame, 35 (1979), pp. 139–153.

[8] S. B. Margolis and B. J. Matkowsky, *Nonlinear stability and bifurication in the transition from laminar to turbulent flame propagation*, Comb. Sci. Techn., 34 (1983), pp. 45–77.

[9] G. Sivashinsky, *Diffusional-thermal theory of cellular flames*, Comb. Sci. Techn., 15 (1977), pp. 137–145.

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1172-1189, September 1990
© 1990 Society for Industrial and Applied Mathematics
004

BOUNDEDNESS AND DECAY RESULTS FOR REACTION-DIFFUSION SYSTEMS*

JEFF MORGAN†

Abstract. Boundedness and decay results are obtained for semilinear parabolic systems of partial differential equations. m-component systems of the form

$$u_t = D\triangle u + f(u) \quad \text{on } \Omega \times \mathbb{R}_+$$

with bounded initial data and various boundary conditions are considered, where D is an $m \times m$ positive diagonal matrix, Ω is a smooth bounded domain in $\mathbb{R}^n$, and $f: \mathbb{R}^m \to \mathbb{R}^m$ is locally Lipschitz. These results are based upon f satisfying a Lyapunov-type condition. The theory is applied to some specific reaction-diffusion problems.

Key words. reaction-diffusion, boundedness, decay, Lyapunov function

AMS(MOS) subject classifications. 35K45, 35B35

1. Introduction. In recent work, Hollis, Martin, and Pierre [8] employ a simple Lyapunov-type condition to obtain global existence and boundedness results for semilinear parabolic systems of the form

$$u_{i_t} = d_i\triangle u_i + f_i(u) \quad \text{on } \Omega \times (0, \infty), \quad i = 1, 2, \tag{1.1}$$

with bounded nonnegative initial data and various boundary conditions. (Here $d_1, d_2 > 0$, $\Omega \in \mathbb{R}^n$ is a smooth bounded domain, and $f = (f_i)$ is locally Lipschitz and of polynomial type.) They impose conditions on f to guarantee that u_1 and u_2 remain nonnegative, and they assume u_1 to be uniformly bounded. Their Lyapunov-type condition is given as

$$f_1(v_1, v_2) + f_2(v_1, v_2) \leqq N \quad \text{for all } v_1, v_2 \geqq 0, \tag{1.2}$$

with the size of the constant N dependent only upon the type of boundary condition. Based on these assumptions, they have been able to prove global existence and boundedness results for (1.1).

More recently, Morgan [13] has extended the global existence portion of these results to m-component systems of the form

$$u_{i_t} = d_i\triangle u_i + f_i(u), \qquad \Omega \times (0, \infty), \quad i = 1, \cdots, m. \tag{1.3}$$

Morgan assumes the existence of an invariant region M of $\mathbb{R}^m$ (such as $\mathbb{R}^m_+$) and a smooth, functional H: $M \to \mathbb{R}$ (deemed a Lyapunov-type function) of the form $H(u) = \sum_{i=1}^m h_i(u_i)$, with h_i convex and nonnegative, such that

$$\partial H(v) f(v) \leqq N \quad \text{for all } v \in M. \tag{1.4}$$

Condition (1.4) is used to obtain $L^2(\Omega \times (0, t))$ bounds on $H(u)$. These bounds are then combined with a so-called "intermediate sums" condition and polynomial growth restrictions to obtain global existence results for (1.3). In this present work, we extend these ideas to generalize the boundedness results of Hollis, Martin, and Pierre, and to obtain decay results. Our primary tool in obtaining these results is an extension of a duality technique that originated in [8].

* Received by the editors December 19, 1988; accepted for publication (in revised form) October 2, 1989. This work was supported in part by the National Science Foundation under grant DMS-8813071.

† Department of Mathematics, Texas A&M University, College Station, Texas 77843.

We should note that our "intermediate sums" condition is motivated by Rothe [16]. Rothe develops estimates for solutions of scalar equations such as

$$u_t = d\triangle u + f(u), \qquad \Omega\times(0,\infty),$$

subject to various boundary conditions and initial data. His primary assumption is that $|f(u)|\leqq K(|u|^r+1)$ and that some $L^a(\Omega\times(0,T))$ a priori bound is known for u. He then gives conditions relating a and r which guarantee both global existence and certain growth estimates for u. In our work, we employ the structure of H to obtain $v=(v_i)=(h_i(u_i))$ as a "subsolution" of

$$v_{i_t}\leqq d_i\triangle v_i + g_i(v), \qquad \Omega\times(0,\infty),$$

where $g_i(v)=h_i'(u_i)f_i(u)$. The "intermediate sums" condition then requires the existence of $A=(a_{ij})\in\mathbb{R}^{m\times m}$ lower triangular satisfying $a_{ij}\geqq 0$ and $a_{ii}>0$ for all $1\leqq i,j\leqq m$ such that $\sum_{i=1}^{j} a_{ji}g_i(v)\leqq K(H(u)^r+1)$ for all $1\leqq j\leqq m$. We then employ (1.4) to obtain an $L^a(\Omega\times(0,T))$ a priori bound for v and give conditions relating a and r guaranteeing both global existence and certain growth estimates for v. These estimates yield similar estimates for u.

We remark that similar problems have been solved in Kanel [9] and conceptually related work can be found in Bates [1], Redheffer, Redlinger, and Walter [15] and Weinberger [20]. Thoroughgoing mathematical surveys of reaction-diffusion systems are given by Rothe [16] and Smoller [18].

Our work is organized as follows. In § 2 we develop notation and state our main results. In § 3 we establish some preliminary estimates, and in § 4 we prove our main results. In § 5 we develop some a priori estimates, and in § 6 we apply our results to some specific reaction-diffusion problems.

2. Notation and statements of results. We assume that the standard L^p and Sobolev spaces are familiar to the reader. If $0\leqq\tau<T$ and $1\leqq p<\infty$, then $W^{2,1,p}(\Omega\times(\tau,T))$ will denote the Banach space consisting of the elements u of $L^p(\Omega\times(\tau,T))$, having the distributional derivatives $\partial_t^r\partial_x^s u$, where $2r+s\leqq 2$ and each of the derivatives lie in $L^p(\Omega\times(\tau,T))$. The norm is defined by

$$\|u\|^{(2)}_{p,\Omega\times(\tau,T)} = \sum_{2r+s\leqq 2} \|\partial_t^r\partial_x^s\|_{p,\Omega\times(\tau,T)}.$$

The positive orthant of $\mathbb{R}^m$ is defined by $\mathbb{R}^m_+=\{x\in\mathbb{R}^m \mid x_i\geqq 0 \text{ for all } 1\leqq i\leqq m\}$.

Throughout, Ω will be a bounded domain in $\mathbb{R}^n$ with smooth boundary $\partial\Omega$ (say $\partial\Omega$ is an $n-1$ dimensional $C^{2+\mu}$ manifold such that Ω lies locally on one side of $\partial\Omega$). Furthermore, if s and t are real numbers satisfying $0\leqq s\leqq t$, then $Q(s,t)$ will denote $\Omega\times(s,t)$. In addition, $Q(0,\infty)$ will denote $\Omega\times(0,\infty)$.

The primary concern of this work is the system

$$\begin{aligned} &u_t(x,t) = D\triangle u(x,t)+f(u(x,t)), && x\in\Omega, \quad t>0,\\ &Bu(x,t)=0, && x\in\partial\Omega,\ t>0,\\ &u(x,0)=u_0(x), && x\in\Omega, \end{aligned} \tag{2.1}$$

where $D=\operatorname{diag}\{d_1,\cdots,d_m\}$ is positive, $f=(f_i)\colon\mathbb{R}^m\to\mathbb{R}^m$ is locally Lipschitz, $\triangle$ is the Laplacian operator, and B and u_0 satisfy the following assumptions.

(A1) $B=(B_i)$ is a diagonal operator given by $Bu=(B_iu_i)$, where $B_iu_i=\alpha_iu_i+\beta(\partial u_i/\partial\eta)$ for all $1\leqq i\leqq m$.

Here $\alpha=(\alpha_i)\in\mathbb{R}^m_+$ and $\beta\in\{0,1\}$ satisfy:

(i) If $\beta=0$ then $\alpha_k=1$ for all $1\leqq k\leqq m$;

(ii) If $\alpha_i=0$ for some $1\leqq i\leqq m$, then $\alpha\equiv 0$ and $\beta=1$.

(A2) $u_0=(u_{0_i})\in L^\infty(\Omega,\mathbb{R}^m)$.

The notation $\partial/\partial\eta$ above represents the derivative with respect to the outward unit normal on $\partial\Omega$. In all that follows, ∇ is the gradient operator and ∂ is the derivative operator.

Conditions (A1), (A2) guarantee local existence and uniqueness for (2.1). A proof of the following result can be found in [8, Prop. 1, p. 745].

PROPOSITION 2.2. *Suppose* (A1), (A2) *hold. Then there exists* $T_{\max}>0$ *and* $N=(N_i)\in C([0, T_{\max}), \mathbb{R}^m)$ *such that*

(i) *System* (2.1) *has a unique, classical, noncontinuable solution* $u(x, t)$ *on* $\bar{\Omega}\times[0, T_{\max})$;

(ii) $\|u_i(\cdot, t)\|_{\infty,\Omega}\leqq N_i(t)$ *for all* $1\leqq i\leqq m, 0\leqq t<T_{\max}$.

Moreover, if $T_{\max}<\infty$, then $\|u_i(\cdot, t)\|_{\infty,\Omega}\to\infty$ for some $1\leqq i\leqq m$.

For the remainder of this paper $\mathbf{M}$ will be an unbounded region of $\mathbb{R}^m$ for which (2.1) is invariant. Since we are primarily interested in the case when the d_i are distinct, we are motivated by Chueh, Conley, and Smoller [3] to assume there exist (possibly unbounded) intervals $\mathbf{M}_i$ of $\mathbb{R}$ such that $\mathbf{M}=\mathbf{M}_1\times\cdots\times\mathbf{M}_m$. In addition, to accommodate this invariance assumption, we require:

(A3) If $1\leqq i\leqq m$ then $u_{0_i}(x)\in\mathbf{M}_i$ for all $x\in\Omega$.

We are now ready to postulate the existence of a symmetric generalized Lyapunov-type structure for (2.1). Namely, we assume there exist $H\in C^2(\mathbf{M}, \mathbb{R}_+)$ and $h_i\in C^2(\mathbf{M}_i, \mathbb{R}_+)$ for $1\leqq i\leqq m$ such that

(H1) $H(z)=\sum_{i=1}^m h_i(z_i)$ for all $z\in\mathbf{M}$;

(H2) $H(z)=0$ if and only if $z=(0, \cdots, 0)^T$;

(H3) $\partial^2 H(z)$ is nonnegative for all $z\in M$;

(H4) there exist $L_1, L_2\in\mathbb{R}$ such that $\partial H(z)f(z)\leqq L_1H(z)+L_2$ for all $z\in M$.

Conditions (H1)-(H3) have some immediate straightforward consequences that we will need in the work below.

PROPOSITION 2.3. *Suppose* u *solves* (2.1) *on* $\bar{\Omega}\times[0, T_{\max})$ *and* (A1)-(A3), (H1)-(H3) *are satisfied. Then*

(i) $h_i(z_i), h_i''(z_i)\geqq 0$ *for all* $z_i\in M_i, 1\leqq i\leqq m$, *and* $h_i(z_i)=0$ *if and only if* $z_i=0$;

(ii) $h_i'(z_i)\neq 0$ *if* $z_i\neq 0$ *for all* $1\leqq i\leqq m$;

(iii) $H(z)\to\infty$ *as* $|z|\to\infty$ *in* M;

(iv) $\alpha_i h_i(u_i(x, t))+\beta\partial(h_i(u_i(x, t)))/\partial\eta\leqq 0$ *for all* $(x, t)\in\partial\Omega\times(0, T_{\max})$ *and* $1\leqq i\leqq m$.

Before we can state boundedness and decay results for (2.1), we need to know that solutions of (2.1) exist globally. That is, we need $T_{\max}=\infty$. The following assumptions are very helpful in this respect.

(H5) There exists $A=(a_{ij})\in\mathbb{R}^{m\times m}$ satisfying $a_{ij}\geqq 0$ and $a_{ii}>0$ for all $1\leqq i, j\leqq m$, such that for each $1\leqq j\leqq m$ either

(i) There exist $r, K_1, K_2\geqq 0$ independent of j such that

$$\sum_{i=1}^{j} a_{ji}h_i'(z_i)f_i(z)\leqq K_1(H(z))^r+K_2$$

for all $z\in M$, or

(ii) There exists $\tilde{p}\geqq 1$ such that for all $\tilde{p}\leqq p<\infty$ there exist $0<\delta_p<1$ and $K_{3p}, K_{4p}\in C(\mathbb{R}_+^2, \mathbb{R}_+)$ such that for all $0\leqq\tau<T\leqq T_{\max}$

$$\|h_j(u_j)\|_{p,Q(\tau,T)}\leqq K_{3p}(\tau, T)+K_{4p}(\tau, T)\|H(u)\|_{p,Q(\tau,T)}^{\delta_p}.$$

(H6) There exists $q_1, K_5, K_6\geqq 0$ such that for all $1\leqq i\leqq m$ and $z\in M$

$$h_i'(z_i)f_i(z)\leqq K_5(H(z))^{q_1}+K_6.$$

We can now state a global existence result. The proof can be found in Morgan [13].

PROPOSITION 2.4. *Suppose* (A1)-(A3) *and* (H1)-(H3), (H5), *and* (H6) *hold. If there exist* $a>0$ *and* $g\in C([0,\infty))$ *such that*

$$\left(\int_0^t\int_\Omega (H(u(x,s)))^a\,dx\,ds\right)^{1/a}\leqq g(t)$$

for all $0<t<T_{\max}$ *and* $r<1+2a/(n+2)$, *then* $T_{\max}=\infty$.

Some remarks are probably in order at this point. First, condition (H5) states that there is either some cancelling of higher-order terms for the jth "intermediate sum" or $h_j(u_j)$ satisfies an L^p growth restriction. Clearly (H5(ii)) is satisfied if u_j can be bounded a priori on $Q(0,T_{\max})$. Second, note that Proposition 2.4 is dependent upon our being able to obtain an L^a bound for $H(u)$ and then comparing this with the size of r. We show in § 5 that this bound can be obtained with $a=2$ provided (H1)-(H4) hold. Consequently, we are guaranteed global existence provided $r<(n+6)/(n+2)$. This condition is satisfied by many reaction-diffusion models (see § 6).

We are now ready to state our boundedness result. Since the hypotheses of the results below contain those of Proposition 2.4, we assume throughout that $T_{\max}=\infty$.

THEOREM 2.5. *Suppose the hypotheses of Proposition* 2.4 *hold with* $\tilde{p}=1$, *and for all* $0<\tau<T$ *and* $p\geqq 1$

$$K_{3p}(\tau,T),\ K_{4p}(\tau,T),\ \|H(u)\|_{a,Q(\tau,T)}\leqq g(T-\tau). \tag{2.6}$$

If $1\leqq r<a$ *or* $1\leqq r=a<(n+2)/n$, *then there exists* $N>0$ *such that* $\|u_i(\cdot,t)\|_{\infty,Q(0,\infty)}\leqq N$ *for all* $1\leqq i\leqq m$.

We can also state a decay result.

THEOREM 2.7. *Suppose the hypotheses of Theorem* 2.5 *hold and* $K_2=K_6=0$. *If there exists* $T>0$ *such that for all* $p\geqq 1$

$$\lim_{t\to\infty}[\|H(u)\|_{a,Q(t,t+T)}+K_{3p}(t,t+T)+K_{4p}(t,t+T)]=0, \tag{2.8}$$

then $\lim_{t\to\infty}\||u(\cdot,t)|\|_{\infty,\Omega}=0$.

We remark that if (H5(i)) is satisfied for all $1\leqq j\leqq m$, then the conditions on K_{3p} and K_{4p} in Theorems 2.5 and 2.7 are vacuous. We give various conditions in § 5 guaranteeing that (2.6) and (2.8) are satisfied for each of $a=1,2$.

So far, we have not stated a result based solely upon the hypotheses of Proposition 2.4. Our last result comes close to achieving this.

THEOREM 2.9. *Suppose* (A1)-(A3), (H1)-(H3), (H5(i)) *for all* $1\leqq j\leqq m$, *and* (H6) *hold. If* $K_2=K_6=0$ *and there exist* $a, K>0$ *such that*

$$\|H(u)\|_{a,Q(0,\infty)}\leqq K, \tag{2.10}$$

$2a/(n+2)<r<1+2a/(n+2)$ *and* $2a<n+2$ *when* $|\alpha|=0$, *then*

$$\lim_{t\to\infty}\||u(\cdot,t)|\|_{\infty,\Omega}=0.$$

In a forthcoming paper we will examine the asymptotic behavior of solutions to (2.1) under the hypotheses of Theorem 2.5.

3. Preliminary estimates. We follow [8] in developing the following notation. For each $1<p<\infty$ and $j=\{1,\cdots,m\}$ we define $A_{j,p}$ on $L^p(\Omega)$ by

$$\begin{aligned} &A_{j,p}w=d_j\triangle w \quad \text{for } w\in\mathscr{D}(A_{j,p}),\\ &\mathscr{D}(A_{j,p})=\left\{w\in W^{2,p}(\Omega):\ \alpha_j w+\beta\frac{\partial w}{\partial\eta}=0 \text{ on } \partial\Omega\right\}, \end{aligned} \tag{3.1}$$

where α_j, β, and d_j are given in § 2. It is well known (cf. Pazy [14]) that $A_{j,p}$ generates a compact, analytic semigroup $T_{j,p} = \{T_{j,p}(t): t \geqq 0\}$ of bounded linear operators on $L^p(\Omega)$, and that

$$\|T_{j,p}(t)w\|_{p,\Omega} \leqq e^{\lambda t}\|w\|_{p,\Omega} \quad \text{for } t \geqq 0, \quad w \in L^p(\Omega), \tag{3.2}$$

where $\lambda_0 < 0$ if $\alpha_j > 0$ and $\lambda_0 = 0$ if $\alpha_j = 0$. Furthermore, for each $\gamma > 0$, $\lambda > \lambda_0$ the fractional powers $(\lambda I - A_{j,p})^{-\gamma}$ exist and are injective, bounded linear operators on $L^p(\Omega)$. In addition, if for $\gamma > 0$ we define $B_{j,p}^{-\gamma} = (-A_{j,p})^{-\gamma}$ if $\alpha_j > 0$, $B_{j,p}^{-\gamma} = (I - A_{j,p})^{-\gamma}$ if $\alpha_j = 0$, and $B_{j,p}^{\gamma} = (B_{j,p}^{-\gamma})^{-1}$, then $\mathcal{D}(B_{j,p}^{\gamma})$ is a Banach space with graph norm $|w|_{\gamma} = \|B_{j,p}^{\gamma}\|_{p,\Omega}$. Also, for $\gamma_1 > \gamma_2 \geqq 0$, $\mathcal{D}(B_{j,p}^{\gamma_1})$ is a dense subspace of $\mathcal{D}(B_{j,p}^{\gamma_2})$ with the inclusion $\mathcal{D}(B_{j,p}^{\gamma_1}) \subset \mathcal{D}(B_{j,p}^{\gamma_2})$ compact (here $\mathcal{D}(B_{j,p}^{0}) = L^p(\Omega)$). The following lemma is quite useful.

LEMMA 3.3. *Suppose $T_{j,p}$ and $B_{j,p}^{\gamma}$ are as above. Then*

(i) $T_{j,p}(t): L^p(\Omega) \to \mathcal{D}(B_{j,p}^{\gamma})$ *for all* $t > 0$;

(ii) $\|B_{j,p}^{\gamma} T_{j,p}(t)w\|_{p,\Omega} \leqq C_{\gamma,p} t^{-\gamma} e^{\lambda_0 t}\|w\|_{p,\Omega}$ *for* $t > 0$ *and* $w \in L^p(\Omega)$;

(iii) $T_{j,p}(t)B_{j,p}^{\gamma} w = B_{j,p}^{\gamma} T_{j,p}(t)w$ *for all* $t > 0$, $w \in \mathcal{D}(B_{j,p}^{\gamma})$;

(iv) *If* $\gamma > n/(2p)$ *then* $\mathcal{D}(B_{j,p}^{\gamma}) \subset L^{\infty}(\Omega)$ *and* $\|w\|_{\infty,\Omega} \leqq M_{\gamma,p}\|B_{j,p}^{\gamma} w\|_{p,\Omega}$ *for all* $w \in \mathcal{D}(B_{j,p}^{\gamma})$;

(v) *If* $\mu > n(q-p)/(2pq) \geqq 0$ *then* $\mathcal{D}(B_{j,p}^{\mu}) \subseteq L^q(\Omega)$ *and* $\|w\|_{q,\Omega} \leqq N_{\mu,q}\|B_{j,p}^{\mu} w\|_{p,\Omega}$ *for all* $w \in \mathcal{D}(B_{j,p}^{\mu})$.

The proofs of parts (i)-(iii) and (v) can be found in Pazy [14, p. 74], and the proof of (iv) is contained in Henry [6, p. 40].

In order to rewrite (2.1) as an integral equation system via variation of constants, we define F_i on $(L^p(\Omega))^m$ by

$$[F_i(w)](x) = f_i(w(x)) \quad \text{for } x \in \Omega, \quad w \in L^p(\Omega, \mathbb{R}^m), \tag{3.4}$$

for all $1 \leqq i \leqq m$. Then, by variation of constants (see Pazy [14]) it follows that $u = (u_i)$ is a solution in $(L^p(\Omega))^m$ to the system

$$u_i(t) = T_{i,p}(t)(u_{0_i}) + \int_0^t T_{i,p}(t-s)F_i(u(s))\,ds \tag{3.5}$$

if and only if $u(x, t) = [u(t)](x)$ solves (2.1).

Before continuing, we need some results concerning scalar equations. Let $0 \leqq \tau < T$, $1 < p < \infty$, and $\theta \in L^p(Q(\tau, T))$ such that $\theta \geqq 0$ and $\|\theta\|_{p,Q(\tau,T)} = 1$. For all $1 \leqq i \leqq m$ consider

$$\begin{aligned} &\phi_{i_t} = d_i \triangle \phi_i - \phi_i + \theta \quad \text{on } Q(\tau, T), \\ &b_1 \phi_i + \beta \frac{\partial b}{\partial \eta} = 0 \quad \text{on } \partial\Omega \times (\tau, T), \\ &\phi_i(x, 0) = 0 \quad \text{on } \Omega, \end{aligned} \tag{3.6}$$

where $b_1 = \min\{\alpha_1, \cdots, \alpha_m\}$.

LEMMA 3.7. *For all $1 \leqq i \leqq m$ there exists a unique solution $\phi_i \in W^{2,1,p}(Q(\tau, T))$ for* (3.6). *Furthermore,*

(i) $\phi_i \geqq 0$;

(ii) *There exists $C_p > 0$ such that*

$$\|\phi_i\|_{p,Q(\tau,T)}^{(2)} \leqq C_p.$$

Proof. Part (i) is a consequence of maximum principles. The proof of part (ii) can be found in Hollis [7, Lemma 2.6] or Hollis, Martin, and Pierre [8, Lemma 3, par. 2, p. 759].

We can actually improve on a portion of part (ii) above. We state this result along with an imbedding result.

LEMMA 3.8. *Let* $1<p<\infty$, $1\leqq i\leqq m$ *and suppose* ϕ_i *solves* (3.6).

(i) *If* $1<p<(n+2)/2$ *and* $p\leqq q\leqq p(n+2)/(n+2-2p)$, *then there exists* $\tilde{C}_p>0$ *such that*

$$\|\phi_i\|_{q,Q(\tau,T)}\leqq \tilde{C}_p;$$

(ii) *If* $1<p<\infty$ *then* $\|\phi_i(\cdot,T)\|_{p,\Omega}\leqq p^{1/p}$;

(iii) *If* $1<p<\infty$ *and* $b_1=\min\{\alpha_1,\cdots,\alpha_m\}\neq 0$, *then there exists* $N>0$ *such that*

$$\|\phi_i\|_{p,Q(\tau,T)}\leqq \frac{N}{N+d_i(p-1)}\left[1-\exp\left(-(T-\tau)\frac{N+d_i(p-1)}{N}\right)\right],$$

and if $b_1=0$ *then* $\|\phi_i\|_{p,Q(\tau,T)}\leqq(1-\exp(-(T-\tau)))$.

Proof. Part (i) is a consequence of Lemma 3.10, part (ii) and Ladyzhenskaja [10, p. 80]. We note that it is sufficient to prove parts (ii) and (iii) for θ smooth. Let $\varepsilon>0$. Then from (3.9) we have

$$(\phi_i+\varepsilon)^{p-1}\phi_{i_t}-d_i(\phi_i+\varepsilon)^{p-1}\triangle\phi_i+(\phi_i+\varepsilon)^{p-1}\phi_i=(\phi_i+\varepsilon)^{p-1}\theta.$$

Consequently, if we integrate over $Q(\tau,t)$ for $\tau\leqq t\leqq T$ and let $\varepsilon\to 0^+$, then we obtain

$$\frac{1}{p}\int_\Omega(\phi_i(x,t))^p\,dx+d_ib_1\int_\tau^t\int_{\partial\Omega}\phi_i^p\,d\sigma\,ds+d_i(p-1)\int_\tau^t\int_\Omega\phi_i^{p-2}|\nabla\phi_i|^2\,dx\,ds$$
$$+\int_\tau^t\int_\Omega\phi_i^p\,dx\,ds\leqq\left(\int_\tau^t\int_\Omega\phi_i^p\,dx\,ds\right)^{1-(1/p)}.$$

Now, $\int_\tau^t\int_\Omega\phi_i^{p-2}|\nabla\phi_i|^2\,dx\,ds=(4/p^2)\int_\tau^t\int_\Omega|\nabla(\phi_i^{p/2})|^2\,dx\,ds$, and if $b_1\neq 0$ then there exists $N>0$ such that

$$\int_\tau^t\int_\Omega\phi_i^p\,dx\,dt\leqq\frac{4N}{p^2}\int_\tau^t\int_\Omega|\nabla(\phi_i^{p/2})|^2\,dx\,ds+\frac{Nb_1}{p-1}\int_\tau^t\int_{\partial\Omega}\phi_i^p\,d\sigma\,ds.$$

If we substitute this above and set $w(t)=\int_\tau^t\int_\Omega\phi_i^p\,dx\,ds$, then we obtain

$$w'(t)+p\left(1+\frac{d_i(p-1)}{N}\right)w(t)\leqq p(w(t))^{1-1/p},\qquad \tau\leqq t\leqq T$$

and $w(\tau)=0$.

A straightforward calculation yields part (iii) for $b_1\neq 0$. Parts (ii) and (iii) for $b_1=0$ are trivial consequences of the integral inequality above.

In all that follows, we set $\hat{\theta}(x,t)=\theta(x,T+\tau-t)$ and $\psi_i(x,t)=\phi_i(x,T+\tau-t)$ on $\bar{\Omega}\times[\tau,T]$. We note that ψ_i is the unique solution of

$$\begin{aligned}&\psi_{i_t}=-d_i\triangle\psi_i+\psi_i-\hat{\theta}\quad\text{on } Q(\tau,T),\\ &b_1\psi_i+\frac{\partial\psi_i}{\partial\eta}=0\quad\text{on } \partial\Omega\times(\tau,T),\\ &\psi_i(\cdot,T)=0\quad\text{on }\Omega.\end{aligned}\tag{3.9}$$

The next two lemmas combine with Lemmas 3.7 and 3.8 to give technical results needed in the proofs of Theorems 2.5 and 2.7.

LEMMA 3.10. *Suppose* ψ_i *solves* (3.9).

(i) *If* $p>(n+2)/2$ *then there exists* $K_{p(T-\tau)}>0$ *such that* $\|\psi_i\|_{\infty,Q(\tau,T)}\leqq K_{p(T-\tau)}$.

(ii) *If* $1<p<(n+2)/2$ *and* $1<q<np/(n-2(p-1))$ *then there exists* $K_{p,q,(T-\tau)}>0$ *such that*

$$\|\psi_i(\cdot,\tau)\|_{q,\Omega}\leqq K_{p,q,(T-\tau)}.$$

Proof. For part (ii), our hypothesis implies there exists $0<\mu<1$ such that $\mu>(n(q-p))/2pq$ and $\mu p/(p-1)<1$. Thus from Lemma 3.3 and the definition of ψ_i we have

$$\begin{aligned}\|\psi_i(\cdot,\tau)\|_{q,\Omega} &\leqq N_{\mu q}\|B^{\mu}_{i,p}\psi_i(\cdot,\tau)\|_{p,\Omega}\\ &\leqq N_{\mu q}C_{\mu p}\int_0^{T-\tau}(T-\tau-s)^{-\mu}\|\theta(\cdot,T-s)\|_{p,\Omega}\,ds,\end{aligned}\tag{3.11}$$

and from Hölder's inequality we obtain

$$\int_0^{T-\tau}(T-\tau-s)^{-\mu}\|\theta(\cdot,T-s)\|_{p,\Omega}\,ds\leqq\left[\frac{(T-\tau)^{1}-\mu p/(p-1)}{1-\mu p/(p-1)}\right]^{(p-1)/p}.\tag{3.12}$$

Combining (3.11) and (3.12) proves part (ii). For part (i), note that if $2p>n+2$ then the choice $q=\infty$ is admissible.

LEMMA 3.13. *If $1\leqq r<k$ and there exists $0<\mu<2$ such that $r+(2-\mu)/(n+2)<1+2k/(n+2)$, then there exist $\delta>1$ and $1<p<(n+2)/2$ such that*

(i) $$k=\frac{np}{(n+2)(p-1)}\delta,$$

(ii) $$\frac{k}{k-r}\leqq\frac{p(n+2)}{n+2-2p},$$

(iii) $$\frac{p}{p-1}\geqq\frac{n+2}{n+\mu}k.$$

Proof. Since $1\leqq r<k$, there exists $0<\varepsilon<k-1$ such that $[(n+2)/2]r\leqq k+(n/2)(k-\varepsilon)$. Set $\delta=\min\{(n+\mu)/n,\,k-\varepsilon\}$ and $p=(n+2)k/((n+2)k-n\delta)$. Then $1<p<(n+2)/2$ and $\delta>1$. A simple calculation shows that (i)-(iii) hold.

The final lemma in this section is a straightforward technical result used in the proof of Theorem 2.9. We omit the tedious proof.

LEMMA 3.14. *Suppose $a, r>0$, $n\in\mathbb{N}$, and $2a/(n+2)<r<1+2a/(n+2)$. Then there exists $\varepsilon>0$ such that if $1<p<1+\varepsilon$ and*

$$k=\frac{ap[(p-1)(n+2)+2p]-arp(p-1)(n+2)}{[p-a(p-1)][(p-1)(n+2)+2p]},$$

then

(i) $$\frac{a}{k}>1,$$

(ii) $$\left(\frac{rp(n+2)}{(p-1)(n+2)+2p}-k\right)\left(\frac{a}{a-k}\right)=\frac{p}{p-1},$$

(iii) $$\left(\frac{p-1}{p}+\frac{2}{n+2}\right)\left(\frac{a-k}{a}\right)<\frac{p-1}{p}.$$

4. Proofs of Theorems 2.5, 2.7, and 2.9. In this section we combine the estimates in § 3 with the hypotheses of Theorems 2.5, 2.7, and 2.9 to obtain our results.

The following lemma is critical to the proofs of Theorems 2.5, 2.7, and 2.9.

LEMMA 4.1. *Suppose that* (H1)-(H3) *and* (H6) *hold, $p>(n+2)/2$ and there exist $T>0$, a sequence $\{t_i\}$, and $\tilde{g}\in C([0,\infty))$ such that*

(i) *$t_1>0$ and $T/4<t_{i+1}-t_i<T/2$ for all $i\geqq 1$;*

(ii) *$\|H(u)\|_{p,Q(\tau,\tau+T)}$, $\|(H(u))^{q_i}\|_{p,Q(\tau,\tau+T)}\leqq\tilde{g}(\tau)$ for all $\tau\geqq 0$;*

(iii) *$\|h_j(u_j(\cdot,t_i))\|_{p,\Omega}\leqq\tilde{g}(t_i)$ for all $i\geqq 1$, $1\leqq j\leqq m$.*

If there exists $K>0$ such that $\tilde{g}(t)\leqq K$ for all $t\geqq 0$, then there exists $N>0$ such that $\|u_i(\cdot,t)\|_{\infty,\Omega}\leqq N$ for all $t\geqq 0, 1\leqq i\leqq m$. Furthermore, if $K_6=0$ and $\lim_{t\to\infty}\tilde{g}(t)=0$, then $\lim_{t\to\infty}\|u(\cdot,t)\|_{\infty,\Omega}=0$.

Proof. Suppose $t>t_5$. Then there exists $i\geqq 1$ such that $t_i<t_{i+1}<t<t_{i+2}$. Note that $t-t_i<T$. Also, there exists $\delta>0$ such that $p=(n+2+\delta)/2$. Hence, if $\mu=n/(n+2)$, then $n/2p<\mu<1$ and $p\mu/(p-1)<1$. Now, note that for each $1\leqq i\leqq m$, $h_i(u_i)$ satisfies

$$(h_i(u_i(x,t)))_t\leqq d_i\triangle(h_i(u_i(x,t)))+K_5(H(u(x,t)))^{q_1}+K_6 \quad \text{on } Q(0,\infty),$$

$$\alpha_i h_i(u_i(x,t))+\beta\frac{\partial(h_i(u_i(x,t)))}{\partial\eta}\leqq 0 \quad \text{on } \partial\Omega\times(0,\infty),$$

$$h_i(u_i(x,0))=h_i(u_{0_i}(x)) \quad \text{on } \Omega.$$

Thus, if we set

$$w_i(\tau)=T_{i,p}(\tau)(h_i(u_{0_i}))+\int_0^\tau T_{i,p}(\tau-s)G_i(s)\,ds$$

for all $\tau>0$ and $1\leqq i\leqq m$, where $G_i(s)=K_5(H(u(\cdot,s)))^{q_1}+K_6$, then maximum principles imply $\|h_i(u_i(\cdot,\tau))\|_{\infty,\Omega}\leqq\|w_i(\tau)\|_{\infty,\Omega}$ for all $\tau>0$. Consequently, from (3.5) and Lemma 3.3,

$$\begin{aligned}\|h_j(u_j(\cdot,t)\|_{\infty,\Omega}&\leqq\|w_j(t)\|_{\infty,\Omega}\\&\leqq M_{\mu\beta}C_{\mu p}[(t-t_i)^{-\mu}\|h_j(u_j(\cdot,t_i))\|_{p,\Omega}\\&\qquad+\|(t-s)^{-\mu}e^{\lambda_0(t-s)}\|_{p/(p-1),(t_i,t)}((K_5+1)\tilde{g}(t_i)+K_6(t-t_i)^{1/p})]\\&\leqq M_{\mu p}C_{\mu p}\left[\left(\frac{T}{4}\right)^{-\mu}\tilde{g}(t_i)+\|(T-s)^{-\mu}e^{\lambda_0((T/4)-s)}\|_{p/(p-1),(0,T/4)}\right.\\&\qquad\left.\cdot((K_5+1)\tilde{g}(t_i)+K_6T^{1/p})\right].\end{aligned}\tag{4.2}$$

Note that since $p\mu/(p-1)<1$, there exists $N_\mu>0$ such that

$$\|(T-s)^{-\mu}e^{\lambda_0((T/4)-s)}\|_{p/(p-1),(0,T/4)}\leqq N_\mu.\tag{4.3}$$

Thus for all $t>t_5$

$$\|h_j(u_j(\cdot,t))\|_{\infty,\Omega}\leqq M_{\mu p}C_{\mu p}\left[\left(\frac{T}{4}\right)^{-\mu}K+N_\mu((K_5+1)K+K_6T^{1/p})\right].$$

Consequently, from Proposition 2.2 (ii) and Proposition 2.3 (i), (iii) there exists $N>0$ such that $\|u_i(\cdot,t)\|_{\infty,\Omega}\leqq N$ for all $1\leqq i\leqq m$, $t\geqq 0$.

Now, suppose $\lim_{t\to\infty}\tilde{g}(t)=0$ and $K_6=0$. Then (4.2) and (4.3) yield

$$\|h_j(u_j(\cdot,t))\|_{\infty,\Omega}\leqq M_{\mu p}C_{\mu p}\left[\left(\frac{T}{4}\right)^{-\mu}+N_\mu(K_5+1)\right]\|\tilde{g}\|_{\infty,(t-T,t)}$$

for all $t>t_5$. Therefore, $\lim_{t\to\infty}\|h_j(u_j(\cdot,t))\|_{\infty,\Omega}=0$, and hence (H2) implies $\lim_{t\to\infty}\|u_j(\cdot,t)\|_{\infty,\Omega}=0$.

Proof of Theorem 2.5. Set $M_1=g(3)$. Note that since $\|H(u)\|_{a,Q(\tau,T)}\leqq g(T-\tau)$ for all $0\leqq\tau<T$, there exists a sequence $\{t_{1,i}\}_{i=1}^\infty$ such that $t_{1,1}>0$, $1<t_{1,i+1}-t_{1,i}<3$ and $\|H(u(\cdot,t_{1,i}))\|_{a,\Omega}\leqq M_1$ for all $i\geqq 1$. Thus, if $T>t_{1,1}$, then there exists $i\geqq 1$ such that $t_{1,i}<T\leqq t_{1,i+1}$. Set $\tau=t_{1,i}$.

Case 1. Suppose $1 \leqq r < a$. Since $r < 1 + 2a/(n+2)$, there exists $0 < \mu < 2$ such that $r + (2-\mu)/(n+2) < 1 + 2a/(n+2)$. Then from Lemma 3.13 there exist $\delta > 1$ and $1 < p < (n+2)/2$ such that $a = np\delta/((n+2)(p-1))$, $a/(a-r) \leqq p(n+2)/(n+2-2p)$ and $p/(p-1) \geqq (n+2)/(n+\mu)a$. Set $p_1 = p/(p-1)$. Suppose $1 \leqq j \leqq m$ and there exists $M_2 > 0$ such that for all $l \geqq 1$ and $1 \leqq i < j$, $\|h_i(u_i)\|_{p_1, Q(t_{1,b}, t_{1,l+1})} \leqq M_2$. Now, if $j \leqq k \leqq m$ such that

$$\|h_i(u_i)\|_{p_1, Q(\tau, T)} \leqq K_{3p_1}(\tau, T) + K_{4p_1}(\tau, T)\|H(u)\|^{\delta_{p_1}}_{p_1, Q(\tau, T)} \tag{4.4}$$

is satisfied in the sense of (H5(ii)) for all $j \leqq i < k$, but not for $i = k$, then (H5(i)) holds for k.

Let $\theta \in L^p(Q(\tau, T))$ such that $\theta \geqq 0$ and $\|\theta\|_{p, Q(\tau, T)} = 1$, and suppose ψ_k solves (3.9) with $i = k$. Then for $1 \leqq i \leqq k$, integration by parts, Proposition 2.3, and (3.9) yield

$$\begin{aligned} \int_\tau^T \int_\Omega h_i(u_i)\hat{\theta}\, dx\, dt \leqq (d_i - d_k) \int_\tau^T \int_\Omega h_i(u_i) \triangle \psi_k\, dx\, dt \\ + \int_\Omega h_i(u_i(x, \tau))\psi_k(x, \tau)\, dx \\ + \int_\tau^T \int_\Omega \psi_k [h_i(u_i) + h_i'(u_i) f_i(u)]\, dx\, dt. \end{aligned} \tag{4.5}$$

Thus from (H5(i)) we have

$$\begin{aligned} \int_\tau^T \int_\Omega \sum_{i=1}^{k} a_{k_i} h_i(u_i)\hat{\theta}\, dx\, dt \leqq \sum_{i=1}^{k-1} a_{ki}(d_i - d_k) \int_\tau^T \int_\Omega h_i(u_i) \triangle \psi_k\, dx\, dt \\ + \sum_{i=1}^{k} a_{ki} \int_\Omega h_i(u_i(x, \tau))\psi_k(x, \tau)\, dx \\ + \int_\tau^T \int_\Omega \psi_k [K_1 (H(u))^r + K_2]\, dx\, dt \\ + \int_\tau^T \int_\Omega \psi \sum_{i=1}^{k} a_{ki} h_i(u_i)\, dx\, dt. \end{aligned} \tag{4.6}$$

We consider the right-hand side of (4.6) termwise. For the first term we have

$$\begin{aligned} &\sum_{i=1}^{k-1} a_{ki}(d_i - d_k) \int_\tau^T \int_\Omega h_i(u_i) \triangle \psi_k\, dx\, dt \\ &\quad = \sum_{i=1}^{j-1} a_{ki}(d_i - d_k) \int_\tau^T \int_\Omega h_i(u_i) \triangle \psi_k\, dx\, dt \\ &\qquad + \sum_{i=j}^{k-1} a_{ki}(d_i - d_k) \int_\tau^T \int_\Omega h_i(u_i) \triangle \psi_k\, dx\, dt \\ &\quad \leqq \sum_{i=1}^{j-1} a_{ki} |d_i - d_k| M_2 C_p \\ &\qquad + \sum_{i=1}^{k-1} a_{ki} |d_i - d_k| C_p [K_{3p_1}(\tau, T) + K_{4p_1}(\tau, T)\|H(u)\|^{\delta_{p_1}}_{p_1, Q(\tau, T)}] \\ &\quad \leqq \sum_{i=1}^{k-1} a_{ki} |d_i - d_k| [M_2 + g(3) + g(3)\|H(u)\|^{\delta_{p_1}}_{p_1, Q(\tau, T)}] \end{aligned} \tag{4.7}$$

from Hölder's inequality, Lemma 3.7, and the assumptions above. For the third term we have

$$\int_\tau^T \int_\Omega \psi_k[K_1(H(u))^r + K_2]\,dx\,dt$$

$$\leqq \|\psi_k\|_{a/(a-r),Q(\tau,T)}[K_1\|H(u)\|^r_{a,Q(\tau,T)} + K_2((T-\tau)|\Omega|)^{r/a}] \tag{4.8}$$

$$\leqq \|\psi_k\|_{a/(a-r),Q(\tau,T)}[K_1(g(3))^r + K_2(3|\Omega|)^{r/a}]$$

from Hölder's inequality and the assumptions above. Now, set $q = a/(a-r)$. Then from Hölder's inequality and Lemma 3.8, since $1 \leqq q \leqq p(n+2)/(n+2-2p)$, there exists $C_{p,(T-\tau)} > 0$ such that

$$\|\psi_k\|_{q,Q(\tau,T)} \leqq C_{p,(T-\tau)}. \tag{4.9}$$

Hence if we combine (4.8) and (4.9) then

$$\int_\tau^T \int_\Omega \psi_k[K_1(H(u))^r + K_2]\,dx\,dt \leqq C_{p,(T-\tau)}[K_1(g(3))^r + K_2(3|\Omega|)^{r/a}]. \tag{4.10}$$

The second term can be handled similarly, since $1 < a/(a-1) < np/(n-2(p-1))$ and Lemma 3.10 imply

$$\int_\Omega \psi_k(x,\tau)h_i(u_i(x,\tau))\,dx \leqq \|\psi_k(\cdot,\tau)\|_{a/(a-1),\Omega}\|h_i(u_i(\cdot,\tau))\|_{a,\Omega} \tag{4.11}$$

$$\leqq K_{p,a/(a-1),(T-\tau)}M_1.$$

Finally, for the last term on the right-hand side of (4.6) we have

$$\sum_{i=1}^k a_{ki}\int_\tau^T\int_\Omega \psi_k h_i(u_i)\,dx\,dt \leqq \sum_{i=1}^{k-1} a_{ki}C_p[M_2 + g(3) + g(3)\|H(u)\|^{\delta_{p_1}}_{p_1,Q(\tau,T)}] \tag{4.12}$$

$$+ a_{kk}\int_\tau^T\int_\Omega \psi_k h_k(u_k)\,dx\,dt,$$

in a manner similar to that of (4.7). Also

$$\int_\tau^T\int_\Omega \psi_k h_k(u_k)\,dx\,dt \leqq \|\psi_k\|_{p,Q(\tau,T)}\|h_k(u_k)\|_{p_1,Q(\tau,T)} \tag{4.13}$$

$$\leqq [1-e^{-(T-\tau)}]\|h_k(u_k)\|_{p_1,Q(\tau,T)},$$

from Hölder's inequality and Lemma 3.8. Consequently,

$$\sum_{i=1}^k a_{ki}\int_\tau^T\int_\Omega \psi_k h_i(u_i)\,dx\,dt \leqq \sum_{i=1}^{k-1} a_{ki}C_p[M_2 + g(3) + g(3)\|H(u)\|^{\delta_{p_1}}_{p_1,Q(\tau,T)}] \tag{4.14}$$

$$+ a_{kk}(1-e^{-(T-\tau)})\|h_k(u_k)\|_{p_1,Q(\tau,T)}.$$

Thus if we combine (4.6), (4.7), (4.10), (4.11), and (4.14) we obtain

$$\int_\tau^T\int_\Omega a_{kk}h_k(u_k)\hat{\theta}\,dx\,dt$$

$$\leqq \sum_{i=1}^{k-1} a_{ki}C_p(|d_i - d_k|+1)[M_2 + g(3) + g(3)\|H(u)\|^{\delta_{p_1}}_{p_1,Q(\tau,T)}] \tag{4.15}$$

$$+ K_{p,a/(a-1),(T-\tau)}M_1 + C_{p,(T-\tau)}[K_1(g(3))^r + K_2(3|\Omega|)^{r/a}]$$

$$+ a_{kk}(1-e^{-3})\|h_k(u_k)\|_{p_1,Q(\tau,T)}.$$

Therefore, combining (4.15) and duality, we see that there exist $K_7, K_8 > 0$ such that

$$\|h_k(u_k)\|_{p_1,Q(\tau,T)} \leqq K_7 + K_8 \|H(u)\|^{\delta_{p_1}}_{p_1,Q(\tau,T)} \tag{4.16}$$

for all $\tau = t_{1,i} < T \leqq t_{1,i+1}$ independent of $i \geqq 1$. Hence $K_{3p_1}(\tau, T)$ and $K_{4p_1}(\tau, T)$ can be chosen such that (4.4) is satisfied for k. Contradiction, and therefore (4.4), holds for all $j \leqq i \leqq m$. Thus, since $\|h_i(u_i)\|_{p_1,Q(\tau,T)} \leqq M_2$ for all $1 \leqq i < j$, we have

$$\|H(u)\|_{p_1,Q(\tau,T)} \leqq m(M_2 + K_{3p_1}(\tau, T)) + mK_{4p_1}(\tau, T)\|H(u)\|^{\delta_{p_1}}_{p_1,Q(\tau,T)}. \tag{4.17}$$

Consequently, for all $i \geqq 1$, we have

$$\|H(u)\|_{p_1,Q(t_{1,i},t_{1,i+1})} \leqq \varepsilon^1 m(M_2 + g(3)) + [mg(3)]^{\varepsilon^1}, \tag{4.18}$$

where $\varepsilon^1 = 1/(1-\delta_p)$. That is, we could have chosen M_2 such that $\|h_j(u_j)\|_{p_1,Q(t_{1,i},t_{1,i+1})} \leqq M_2$ for all $1 \leqq j \leqq m$, $i \geqq 1$, and hence such that $\|H(u)\|_{p_1,Q(t_{1,i},t_{1,i+1})} \leqq M_2$ for all $i \geqq 1$. Recall that $p_1 \geqq ((n+2)/(n+\mu))a$, $(n+2)/(n+\mu) > 1$ and $1 < t_{1,i+1} - t_{1,i} < 3$. Consequently, there exists a sequence $\{t_{2,i}\}_{i=1}^{\infty}$ such that $t_{1,2i-1} < t_{2,i} < t_{1,2i}$ for all $i \geqq 1$ and $\|H(u(\cdot, t_{2,i}))\|_{p_1,\Omega} \leqq M_2$ for all $i \geqq 1$. Note that $1 < t_{2,i+1} - t_{2,i} < 9$ for all $i \geqq 1$ and there exists $\tilde{M}_2$ such that $\|H(u)\|_{p_1,Q(t_{2,i},t_{2,i+1})} \leqq \tilde{M}_2$ for all $i \geqq 1$. Now, return to the beginning of Case 1, replace a by p_1, choose the same value of μ, a corresponding value of $p > 1$, and set $p_2 = p/(p-1) \geqq [(n+2)/(n+\mu)]p_1 \geqq [(n+2)/(n+\mu)]^2 a$. Then following the arguments above we find that there exists $M_3 > 0$ such that $\|H(u)\|_{p_2,Q(t_{2,i},t_{2,i+1})} \leqq M_3$ for all $i \geqq 1$.

If we proceed inductively, then for all $k \geqq 2$ there exist $M_{k+1} > 0$, $p_k > [(n+2)/(n+\mu)]^k a$ and a sequence $\{t_{k,i}\}_{i=1}^{\infty}$ such that

(i) $t_{k,1} > 0$ and $1 < t_{k,i+1} - t_{k,i} < 3^k$ for all $i \geqq 1$,

(ii) $\|H(u(\cdot, t_{k,i}))\|_{p_{k-1},\Omega} \leqq M_k$ for all $i \geqq 1$,

(iii) $\|H(u)\|_{p_k,Q(t_{k,i},t_{k,i+1})} \leqq M_{k+1}$ for all $i \geqq 1$.

Note that $\lim_{k\to\infty} p_k = \infty$ since $(n+2)/(n+\mu) > 1$. Hence, by taking k sufficiently large, we can apply the boundedness portion of Lemma 4.1 to obtain $N > 0$ such that $\||u|\|_{\infty,Q[0,\infty)} \leqq N$.

Case 2. Suppose $1 \leqq r = a < (n+2)/n$. Then there exists $0 < \varepsilon < 2$ such that $a < (n+2)/(n+\varepsilon)$. If we set $p = (n+2)/(2-\varepsilon) > (n+2)/2$, let $\theta \in L^p(Q(\tau, T))$ such that $\theta \geqq 0$ and $\|\theta\|_{p,Q(\tau,T)} = 1$, and suppose ψ_k solves (3.9) with $i = k$; then Lemma 3.10 implies $\|\psi_k\|_{\infty,Q(\tau,T)} \leqq K_{p(T-\tau)}$ for all $\tau = t_{1,i} < T \leqq t_{1,i+1}$ independent of $i \geqq 1$. Proceeding as in case 1 with $p_1 = p/(p-1) = (n+2)/(n+\varepsilon)$ we obtain (4.6). The first and last terms on the right-hand side of (4.6) are handled as above. For the second and third terms we have

$$\int_{\Omega} \psi_k(x, \tau) h_i(u_i(x, \tau))\, dx \leqq K_{p(T-\tau)} |\Omega|^{(a-1)/a} M_1, \tag{4.19}$$

$$\int_{\tau}^{T} \int_{\Omega} \psi_k [K_1(H(u))^r + K_2]\, dx\, dt = K_{p(T-\tau)}[K_1(g(3))^a + 3|\Omega|K_2]. \tag{4.20}$$

Substituting (4.7), (4.14), (4.19), and (4.20) into (4.6), and continuing as in case 1, yields $M_2 > 0$ such that $\|H(u)\|_{p_1,Q(t_{1,i},t_{1,i+1})} \leqq M_2$ independent of $i \geqq 1$. Then, since $p_1 > a$, the result follows from case 1.

Proof of Theorem 2.7. We may assume without loss of generality that

$$\lim_{\tau\to\infty} [\|H(u)\|_{a,Q(\tau,\tau+T)} + K_{3p}(\tau, \tau+T) + K_{4p}(\tau, \tau+T)] = 0, \tag{4.21}$$

for all $T > 0$ and $p \geqq 1$. Thus there exist $M_1 \in C([0, \infty))$ and a sequence $\{t_{1,i}\}_{i=1}^{\infty}$ such that $\lim_{t\to\infty} M_1(t) = 0$, $t_1 > 0$, $1 < t_{1,i+1} - t_{1,i} < 3$, and $\|H(u(\cdot, t_{1,i}))\|_{a,\Omega} \leqq M_1(t_{1,i})$ for all $i \geqq 1$.

Consequently, if we proceed as in the proof of Theorem 2.5 with $K_2 = K_6 = 0$ and note that (4.21) holds, then there exists $M_2 \in C([0,\infty))$ such that $\lim_{t\to\infty} M_2(t) = 0$ and $\|H(u)\|_{p_1,Q(t_{1,i},t_{1,i+1})} \leqq M_2(t_{1,i})$ for all $i \geqq 1$, where $p_1 = (n+2)/(n+\mu)$ if $r < a$ and $p_1 = (na+n+2)/2n$ if $1 \leqq r = a < (n+2)/n$. Then, noting that $p_1 > r \geqq 1$, we can proceed as in the proof of Theorem 2.5. That is, for all $k \geqq 2$ there exists $M_{k+1} \in C([0,\infty))$, $p_k \geqq a[(n+2)/(n+\mu)]^{k-1}$ and a sequence $\{t_{k,i}\}_{i=1}^{\infty}$ such that

(i) $\lim_{t\to\infty} M_{k+1}(t) = 0$,

(ii) $t_{k,1} > 0$ and $1 < t_{k,i+1} - t_{k,i} < 3^k$ for all $i \geqq 1$,

(iii) $\|H(u(\cdot, t_{k,i}))\|_{p_{k-1},\Omega} \leqq M_k(t_{k-1,i})$ for all $i \geqq 1$,

(iv) $\|H(u)\|_{p_k,Q(t_{k,i},t_{k,i+1})} \leqq M_{k+1}(t_{k,i})$ for all $i \geqq 1$.

Hence, if we note that $(n+2)/(n+\mu) > 1$, then we can apply the decay portion of Lemma 4.1 to prove that $\lim_{t\to\infty} \||u(\cdot, t)|\|_{\infty,\Omega} = 0$.

Proof of Theorem 2.9. We claim that for all $1 \leqq i \leqq m$ there exists $q_i \geqq 1$ such that for all $q \geqq q_i$ there exists $K_{7q}, K_{8q} > 0$, $0 < \delta_q < 1$ such that

$$\|h_i(u_i)\|_{q,Q(0,T)} \leqq K_{7q} + K_{8q}\|H(u)\|_{q,Q(0,T)}^{\delta q}, \tag{4.22}$$

for all $T > 0$. Let $1 \leqq j \leqq m$ such that (4.22) is satisfied for all $1 \leqq i < j$. Let $\tau = 0$, $T > 0$, and $p > 1$ be given as in Lemma 3.14 such that $p/(p-1) > q_i$ for $1 \leqq i < j$ and $p < (n+2)/2$. Suppose $\theta \in L^p(Q(0,T))$ such that $\theta \geqq 0$ and $\|\theta\|_{p,Q(0,T)} = 1$, and ψ_j is the solution of (3.9) with $i = j$. Then (4.6) becomes

$$\begin{aligned}\int_0^T \int_\Omega \sum_{i=1}^{j} a_{ji} h_i(u_i)\hat{\theta}\, dx\, dt \leqq \sum_{i=1}^{j-1} a_{ji}(d_i - d_j) \int_0^T \int_\Omega h_i(u_i)\triangle\psi_j\, dx\, dt \\ + \sum_{i=1}^{j} a_{ji} \int_\Omega h_i(u_{0_i})\psi_j(x,0)\, dx \\ + \int_0^T \int_\Omega \psi_j K_1 (H(u))^r\, dx\, dt \\ + \int_0^T \int_\Omega \psi_j \sum_{i=1}^{j} a_{ji} h_i(u_i)\, dx\, dt.\end{aligned} \tag{4.23}$$

As in the proof of Theorem 2.5, we consider the right-hand side of (4.23) termwise. Hölder's inequality, Lemma 3.7, and our assumptions above yield the following bound for the first term,

$$\begin{aligned}&\sum_{i=1}^{j-1} a_{ji}(d_i - d_j) \int_0^T \int_\Omega h_i(u_i)\triangle\psi_j\, dx\, dt \\ &\quad \leqq \sum_{i=1}^{j-1} a_{ji}|d_i - d_j| C_p[K_{7p/(p-1)} + K_{8p/(p-1)}\|H(u)\|_{p/(p-1),Q(0,T)}^{\delta p/(p-1)}].\end{aligned} \tag{4.24}$$

For the second term, we apply Hölder's inequality and Lemma 3.8 to obtain

$$\sum_{i=1}^{j} a_{ji} \int_\Omega h_i(u_{0_i})\psi_j(x,0)\, dx \leqq p^{1/p} \sum_{i=1}^{j} a_{ji}\|h_i(u_{0_i})\|_{\infty,\Omega}|\Omega|^{(p-1)/p}. \tag{4.25}$$

We now consider the third and fourth terms. Since $1 < p < (n+2)/2$, if we set $q = (p(n+2))/(n+2-2p)$ then Lemma 3.8 implies $\|\psi_j\|_{q,Q(0,T)} \leqq \tilde{C}_p$. Thus

$$\int_0^T \int_\Omega \psi_j K_1 (H(u))^r\, dx\, dt \leqq K_1 \tilde{C}_p \left[\int_0^T \int_\Omega (H(u))^{p_1}\, dx\, dt\right]^{p_2}, \tag{4.26}$$

where $p1=(rp(n+2))/((p-1)(n+2)+2p)$ and $p2=(p-1)/p+2/(n+2)$. Now, let k be given as in Lemma 3.14. Then $a/k>1$, $(p1-k)(a/(a-k))=p/(p-1)$, and there exists $0<\varepsilon_{p/(p-1)}<1$ such that $p2((a-k)/a)=\varepsilon_{p/(p-1)}((p-1)/p)$. Consequently,

$$\left[\int_0^T\int_\Omega (H(u))^{p1}\,dx\,dt\right]^{p2}=\left[\int_0^T\int_\Omega (H(u))^{p1-k}(H(u))^k\,dx\,dt\right]^{p2} \tag{4.27}$$
$$\leqq K^{kp2}\|H(u)\|^{\varepsilon_{p/(p-1)}}_{p/(p-1),Q(0,T)}.$$

So, combining (4.26) and (4.27) yields

$$\int_0^T\int_\Omega \psi_k K_1(H(u))^r\,dx\,dt\leqq K_1\tilde{C}_p K^{kp2}\|H(u)\|^{\varepsilon_{p/(p-1)}}_{p/(p-1),Q(0,T)}. \tag{4.28}$$

Finally, if $b_1=0$ then $2a/(n+2)<1<1+(2a/(n+2))$ implies that we can handle the fourth term similarly to (4.26)-(4.28). If $b_1\neq 0$, then we can apply (4.22) and Lemma 3.8 to obtain

$$\int_0^T\int_\Omega \psi_j\sum_{i=1}^{j} a_{ji}h_i(u_i)\,dx\,dt=\sum_{i=1}^{j-1}\int_0^T\int_\Omega \psi_j a_{ji}h_i(u_i)\,dx\,dt+\int_0^T\int_\Omega \psi_j a_{jj}h_j(u_j)\,dx\,dt$$
$$\leqq\sum_{i=1}^{j-1} a_{ji}[K_{7p/(p-1)}+K_{8p/(p-1)}\|H(u)\|^{\delta_{p/(p-1)}}_{p/(p-1),Q(0,T)}] \tag{4.29}$$
$$+\frac{Na_{jj}}{N+d_j(p-1)}\|h_j(u_j)\|_{p/(p-1),Q(0,T)}.$$

Consequently, if we apply duality and combine (4.23)-(4.25), (4.28), and (4.29), then we see that there exist $K_{9p/(p-1)}$, $K_{10p/(p-1)}>0$, and $0<\hat{\delta}_{p/(p-1)}<1$ such that

$$\|h_j(u_j)\|_{p/(p-1),Q(0,T)}\leqq K_{9p/(p-1)}+K_{10p/(p-1)}\|H(u)\|^{\hat{\delta}_{p/(p-1)}}_{p/(p-1),Q(0,T)}, \tag{4.30}$$

for all $p/(p-1)$ sufficiently large and $T>0$. Therefore, (4.22) holds for all $1\leqq i\leqq m$. Thus, for all q sufficiently large and $T>0$, we have

$$\|H(u)\|_{q,Q(0,T)}\leqq mK_{7q}+mK_{8q}\|H(u)\|^{\delta q}_{q,Q(0,T)}, \tag{4.31}$$

which implies

$$\|H(u)\|_{q,Q(0,T)}\leqq\frac{mK_{7q}}{1-\delta q}+(mK_{8q})^{1/(1-\delta_q)}, \tag{4.32}$$

for all q sufficiently large and $T>0$. That is, for all q sufficiently large there exist $N_q>0$ such that

$$\|H(u)\|_{q,Q(0,\infty)}\leqq N_q. \tag{4.33}$$

It is now a routine matter to show that the hypotheses of Lemma 4.1 can be satisfied. Our result follows.

5. A priori estimates. In this section we demonstrate that (H1)-(H4) yield certain L^1 and L^2 bounds on $H(u)$, dependent upon the constants L_1, L_2 in (H4) and $b_1=\min\{\alpha_1,\cdots,\alpha_m\}$. These bounds can be combined with Theorems 2.5, 2.7, and 2.9 to yield boundedness and decay results for (2.1). Before beginning, we recall the following result from Morgan [13, Thm. 3.3].

PROPOSITION 5.1. *Suppose* (A1)-(A3) *and* (H1)-(H4) *hold. If u solves* (2.1) *on* $\bar{\Omega}\times[0,T_{\max})$, *and* $T_{\max}<\infty$, *then* $\|H(u)\|_{2,Q(0,T_{\max})}<\infty$.

Consequently, if (H5(i)) is satisfied for all $1 \leqq j \leqq m$ with $r < (n+6)/(n+2)$, then Proposition 2.4 implies $T_{\max} = \infty$. Throughout the remainder of this section we assume that global existence has been established for (2.1) either through Propositions 2.4 and 5.1 or through some other method.

Now, suppose $L_1 \leqq 0$ and $L_2 = 0$ in (H4). Let $1 \leqq k \leqq m$ such that $d_k = \max\{d_1, \cdots, d_m\}$ and let $t_0 \geqq 0$. For $1 \leqq i \leqq m+1$, let z_i be the solution of

$$
\begin{aligned}
&z_{i_t} = d_i \Delta z_i + F_i \quad \text{on } Q(t_0, \infty), \\
&b_1 z_i + \beta \frac{\partial z_i}{\partial \eta} = 0 \quad \text{on } \partial\Omega \times (t_0, \infty), \\
&z_i(x, 0) = M_i \quad \text{on } \Omega,
\end{aligned} \tag{5.2}
$$

where $F_i(x, t) = h_i'(u_i(x, t)) f_i(u(x, t))$ for all $(x, t) \in Q(t_0, \infty)$ and $1 \leqq i \leqq m$, $F_{m+1}(x, t) = -\partial H(u(x, t)) f(u(x, t))$ for all $(x, t) \in Q(t_0, \infty)$, $M_i = h_i(u_i(\cdot, t_0))$ for all $1 \leqq i \leqq m$, $M_{m+1} = 0$ and $d_{m+1} = d_m$. Note that the strong maximum principle implies $z_{m+1} \geqq 0$ and $z_i \geqq h_i(u_i)$ for all $1 \leqq i \leqq m$. Also, $\sum_{i=1}^{m+1} F_i \equiv 0$.

PROPOSITION 5.3. *Suppose* (A1)-(A3) *and* (H1)-(H4) *hold with* $L_1 \leqq 0$ *and* $L_2 = 0$. *Then for all* $0 \leqq \tau < T$, $\|H(u)\|_{1,Q(\tau,T)} \leqq \|H(u_0)\|_{1,\Omega}(T-\tau)$. *If, in addition, either* $L_1 < 0$ *or* $b_1 \neq 0$ *then there exists* $N_1 > 0$ *such that* $\|H(u)\|_{1,Q(0,\infty)} \leqq N_1$.

Proof. Let $t > 0$. From (5.2) with $t_0 = 0$ we have

$$
\|z_i(\cdot, t)\|_{1,\Omega} \leqq \|z_i(\cdot, 0)\|_{1,\Omega} + \int_0^t \int_\Omega F_i(x, s)\, dx\, ds,
$$

for all $1 \leqq i \leqq m$. Consequently, (H4) implies

$$
\left\| \sum_{i=1}^{m+1} z_i(\cdot, t) \right\|_{1,\Omega} \leqq \|H(u_0)\|_{1,\Omega} + L_1 \int_0^t \|H(u(\cdot, s))\|_{1,\Omega}\, dx\, ds. \tag{5.4}
$$

Thus if $L_1 = 0$ then for all $0 \leqq \tau < T$

$$
\|H(u)\|_{1,Q(\tau,T)} \leqq \left\| \sum_{i=1}^{m+1} z_i \right\|_{1,Q(\tau,T)} \leqq \int_\tau^T \|H(u_0)\|_{1,\Omega}\, dt = \|H(u_0)\|_{1,\Omega}(T-\tau). \tag{5.5}
$$

Now, suppose $L_1 < 0$. Then (5.4) implies

$$
\|H(u(\cdot, t))\|_{1,\Omega} \leqq \|H(u_0)\|_{1,\Omega} + L_1 \int_0^t \|H(u(\cdot, s))\|_{1,\Omega}\, ds, \tag{5.6}
$$

and hence

$$
\|H(u)\|_{1,Q(0,\infty)} \leqq -\frac{1}{L_1} \|H(u_0)\|_{1,\Omega}. \tag{5.7}
$$

Finally, suppose $L_1 = 0$ and $b_1 \neq 0$. Let $t_0 > 0$, and for all $t \geqq t_0$ set

$$
w(x, t) = \int_{t_0}^t \sum_{i=1}^{m+1} \frac{d_i}{d_k} z_i(x, s)\, ds.
$$

Then

$$
\begin{aligned}
&w_t = d_k \Delta w + H(u(\cdot, t_0)) + \sum_{i=1}^{m+1} \left(\frac{d_i}{d_k} - 1 \right) z_i \quad \text{on } Q(t_0, \infty), \\
&b_1 w + \beta \frac{\partial w}{\partial \eta} = 0 \quad \text{on } \partial\Omega \times (t_0, \infty), \\
&w(\cdot, t_0) \equiv 0 \quad \text{on } \Omega.
\end{aligned} \tag{5.8}
$$

Hence, if we let v be the unique solution of

$$\begin{aligned} -d_k \triangle v &= H(u(\cdot, t_0)) \quad \text{on } \Omega, \\ b_1 v + \beta \frac{\partial v}{\partial \eta} &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{5.9}$$

then the strong maximum principle implies

$$\|w(\cdot, t)\|_{\infty,\Omega} \leqq \|v\|_{\infty,\Omega} \quad \text{for all } t \geqq t_0. \tag{5.10}$$

Consequently, from the definition of w, the result follows.

PROPOSITION 5.11. *Suppose* (A1)-(A3) *and* (H1)-(H4) *hold with* $L_1 = L_2 = 0$. *If* $b_1 \neq 0$ *then there exists* $N_2 > 0$ *such that* $\|H(u)\|_{2,Q(0,\infty)} \leqq N_2$.

Proof. Suppose $t_0 > 0$ is fixed. Let $T > t_0$. Then from (5.8)

$$\begin{aligned} \int_{t_0}^{T} \int_\Omega z_i \sum_{j=1}^{m+1} \left(\frac{d_j}{d_k} - 1\right) z_j \, dx \, dt &= \int_{t_0}^{T} \int_\Omega z_i [w_t - d_k \triangle w - H(u(x, t_0))] \, dx \, dt \\ &= -\int_{t_0}^{T} \int_\Omega z_i H(u(x, t_0)) \, dx \, dt + \int_{t_0}^{T} \int_\Omega z_i w_t \, dx \, dt \\ &\quad - \frac{d_k}{d_i} \int_{t_0}^{T} \int_\Omega w d_i \triangle z_i \, dx \, dt \\ &= -\int_{t_0}^{T} \int_\Omega z_i H(u(x, t_0)) \, dx \, dt + \int_{t_0}^{T} \int_\Omega z_i w_t \, dx \, dt \\ &\quad - \frac{d_k}{d_i} \int_{t_0}^{T} \int_\Omega w(z_{i_t} - F_i) \, dx \, dt. \end{aligned} \tag{5.12}$$

Consequently,

$$\begin{aligned} \int_{t_0}^{T} \int_\Omega \sum_{i=1}^{m+1} \frac{d_i}{d_k} z_i \sum_{j=1}^{m+1} \left(\frac{d_j}{d_k} - 1\right) z_j \, dx \, dt &= -\sum_{i=1}^{m} \frac{d_i}{d_k} \int_{t_0}^{T} \int_\Omega z_i H(u(x, t_0)) \, dx \, dt \\ &\quad + \sum_{i=1}^{m} \int_\Omega z_i(x, T) w(x, T) \, dx \\ &\quad + \int_{t_0}^{T} \int_\Omega \sum_{i=1}^{m+1} \left(\frac{d_i}{d_k} + 1\right) z_i \sum_{j=1}^{m+1} \frac{d_j}{d_k} z_j \, dx \, dt. \end{aligned} \tag{5.13}$$

Therefore,

$$\begin{aligned} 2 \int_{t_0}^{T} \int_\Omega \sum_{i=1}^{m+1} \frac{d_i}{d_k} z_i^2 \, dx \, dt &\leqq \sum_{i=1}^{m} \frac{d_i}{d_k} \int_{t_0}^{T} \int_\Omega z_i(x, t) H(u(x, t_0)) \, dx \, dt \\ &\leqq \|H(u(\cdot, t_0))\|_{1,\Omega} \|v\|_{\infty,\Omega} \end{aligned} \tag{5.14}$$

from (5.10). The result follows.

We now state one of several possible corollaries to Theorems 2.5, 2.7, and 2.9 and Propositions 5.3 and 5.11.

COROLLARY 5.15. *Suppose* (A1)-(A3), (H1)-(H4), (H5(i)) *for all* $1 \leqq j \leqq m$, *and* (H6) *hold. If* $2/(n+2) < r < (n+6)/(n+2)$, $b_1 \neq 0$ *and* $L_1 = L_2 = K_2 = K_6 = 0$ *then* $\lim_{t \to \infty} \| |u(\cdot, t)| \|_{\infty,\Omega} = 0$.

Before closing this section we remark that the results in Proposition 5.3 and 5.11 can be improved to allow $L_1 < \lambda_0 \min\{d_1, \cdots, d_m\}$, where λ_0 is the principle eigenvalue of $-\triangle$ subject to the boundary conditions $b_1 v + \beta(\partial v / \partial \eta) = 0$.

6. Applications. In this section we give some examples to support our theory. Before we begin, we state a well-known invariance lemma. For a proof see Lightbourne and Martin [12].

LEMMA 6.1. *Suppose* (A1)-(A3) *hold and for all* $1 \leqq i \leqq m, f_i(z) \geqq 0$ *whenever* $z \in \mathbb{R}_+^m$ *and* $z_i = 0$. *Then* $\mathbb{R}_+^m$ *is invariant for* (2.1).

Example 1. First we consider the interaction of oxygen, carbon dioxide, and hemoglobin as blood travels through a pulmonary capillary (cf. [17]). If we let u_1, u_2, u_3, u_4, u_5 represent O_2, HbO_2, CO_2, $HbCO_2$, Hb respectively, then the reaction scheme

$$\mathrm{Hb} + \mathrm{O_2} \rightleftarrows \mathrm{HbO_2}, \qquad \mathrm{Hb} + \mathrm{CO_2} \rightleftarrows \mathrm{HbCO_2},$$

gives rise to the mathematical model (2.1) with

$$f(u) = (f_i(u)) = \begin{cases} k_2u_2 - k_1u_1u_5, \\ -k_2u_2 + k_1u_1u_5, \\ k_4u_4 - k_3u_3u_5, \\ -k_4u_4 + k_3u_3u_5, \\ k_2u_2 + k_4u_4 - k_1u_1u_5 - k_3u_3u_5, \end{cases} \tag{6.2}$$

where $k_1, \cdots, k_5 > 0$ are reaction rates. Clearly, f satisfies the hypotheses of Lemma 6.1 and hence $M = \mathbb{R}_+^5$ is invariant for (6.2). In addition, the choice

$$H(u) = u_1 + 2u_2 + u_3 + 2u_4 + u_5$$

obviously satisfies (H1)-(H4), (H6) with $L_1 = L_2 = K_6 = 0$ and $q_1 = 2$. Furthermore, (H5(i)) is satisfied for all $1 \leqq j \leqq 5$ with $K_1 = \max\{k_2, k_4\}$, $K_2 = 0$, $r = 1$, and

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & \frac{1}{2} & 0 & 0 & 0 \\ 1 & \frac{1}{2} & 1 & 0 & 0 \\ 1 & \frac{1}{2} & 1 & \frac{1}{2} & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Consequently, we can apply Corollary 5.15 to conclude that if $b_1 = \min\{\alpha_1, \cdots, \alpha_5\} > 0$ then $\lim_{t\to\infty} \| |u(\cdot, t)| \|_{\infty,\Omega} = 0$. In the case where $b_1 = 0$, we can apply Proposition 5.3 and Theorem 2.5 to conclude that there exists $N > 0$ such that $\| |u(\cdot, t)| \|_{\infty,\Omega} \leqq N$ for all $t \geqq 0$.

Example 2. Second, we examine a model considered by Lasry [11] (see also Bates and Brown [2]) to study nerve conduction. More specifically, we consider (2.1) with

$$f(u) = \begin{pmatrix} K(1-\rho)u_1 - g(\beta)u_2 \\ g(\beta)u_1 + K(1-\rho)v_2 \end{pmatrix}, \tag{6.3}$$

where $K > 0$, (ρ, β) are polar coordinates for (u_1, u_2), and g is a smooth 2π periodic function. Now, set $M = \mathbb{R}^2$ and

$$H(u) = u_1^2 + u_2^2.$$

Then (H1)-(H4), (H6) are satisfied with $L_1 = K$, $L_2 = 0$, $K_6 = 0$. Furthermore, (H5(i)) is satisfied for $1 \leqq j \leqq 2$ with $K_1 = K + \|g\|_\infty$, $K_2 = 0$, $r = 1$, and

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Now, actually $\partial H(u)f(u) = K[H(u) - (H(u))^{3/2}]$. From this it follows that

$$\frac{d}{dt}\int_\Omega H(u(x,t))\,dx \leqq K\int_\Omega H(u(x,t))\,dx - K|\Omega|^{1/2}\left(\int_\Omega H(u(x,t))\,dx\right)^{3/2}. \tag{6.4}$$

Consequently, (6.4) implies that for all $t \geqq 0$, $\|H(u(\cdot,t))\|_{1,\Omega} \leqq \max\{1/|\Omega|, \|H(u_0)\|_{1,\Omega}\}$. Therefore, we can apply Theorem 2.5 to conclude that there exists $N > 0$ such that $\||u(\cdot,t)|\|_{\infty,\Omega} \leqq N$ for all $t \geqq 0$. More can be said if $b_1 = \min\{\alpha_1, \alpha_2\} \neq 0$. In this case we can improve (6.4) to

$$\begin{aligned}&\frac{d}{dt}\int_\Omega H(u(x,t))\,dx + \lambda_0 \min\{d_1, d_2\}\int_\Omega H(u(x,t))\,dx\\ &\qquad \leqq K\int_\Omega H(u(x,t))\,dx - K|\Omega|^{1/2}\left(\int_\Omega H(u(x,t))\,dx\right)^{3/2},\end{aligned} \tag{6.5}$$

where λ_0 is the principle eigenvalue of $-\triangle$ with boundary conditions $b_1 v + \beta(\partial v/\partial\eta) = 0$. Hence for $\lambda_0 \min\{d_1, d_2\} \geqq K$ we obtain $\lim_{t\to\infty}\|H(u(\cdot,t))\|_{1,\Omega} = 0$. Thus, from Theorem 2.6 we have $\lim_{t\to\infty}\||u(\cdot,t)|\|_{\infty,\Omega} = 0$.

Example 3. Our third example is a system that arises in the modeling of chemical reactions. If we denote by u_1 and u_2 the concentration and the temperature, respectively, of a given reactant, then we might consider (2.1) with

$$f(u) = \begin{pmatrix} -\mu^2 u_1^\rho \exp[\gamma - \gamma/u_2] \\ \nu\mu^2 u_1^\rho \exp[\gamma - \gamma/u_2] \end{pmatrix}, \tag{6.6}$$

for $u_1 \geqq 0$, $u_2 > 0$ and the obvious extension to $u_1 \geqq 0$, $u_2 = 0$. Here, the constants ν and γ are positive and denote the Prater temperature and the Arrehenius number, respectively. μ^2 denotes the Thiele number and $\rho \geqq 1$ is the order of the reaction. (For further information, references, and interesting treatment of a more general model, see Friedman and Tzavares [4].) Clearly Lemma 6.1 is satisfied, and hence $M = \mathbb{R}^2_+$ is invariant for this system. In addition, the choice

$$H(u) = \nu u_1 + u_2$$

satisfies (H1)-(H4), (H6) with $L_1 = L_2 = K_6 = 0$. Furthermore, (H5(i)) is satisfied for $1 \leqq j \leqq 2$ with $K_1 = K_2 = 0$, $r = 1$, and

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Consequently, in the case $b_1 = \min\{\alpha_1, \alpha_2\} = 0$ we can apply Proposition 5.3 and Theorem 2.5 to obtain $N > 0$ such that $\||u(\cdot,t)|\|_{\infty,\Omega} \leqq N$ for all $t \geqq 0$. In the case $b_1 \neq 0$, we can apply Corollary 5.15 to obtain $\lim_{t\to\infty}\||u(\cdot,t)|\|_{\infty,\Omega} = 0$.

Example 4. Our last example in this section is the three-component system (2.1) with

$$f(u) = \begin{pmatrix} -u_1 u_2^2 + \eta_1 u_2^3 \\ u_1 u_2^2 - \eta_1 u_2^3 - \beta(u_2 - \eta_2 u_3) \\ \beta(u_2 - \eta_2 u_3) \end{pmatrix}, \tag{6.7}$$

where β, η_1, $\eta_2 > 0$. f represents the basic kinetics of the Schlögl model due to Gray and Scott [5] (also see Vastano et al. [19]) for the chemical reaction

$$A + 2B \rightleftarrows 3B, \qquad B \rightleftarrows C,$$

we note immediately that Lemma 6.1 yields $M = \mathbb{R}^3_+$ invariant for this model, and that

$$H(u) = u_1 + u_2 + u_3$$

easily satisfies (H1)-(H4), and (H6). However, for this choice of H, the exponent $r=3$ in (H5(i)) is undesirable. Further analysis reveals that (6.7) actually yields a very rich Lyapunov structure. In fact, for all $M \in \mathbb{N}$,

$$H_M(u) = \frac{1}{M}[u_1^M + \eta_1^{M-1} u_2^M + (\eta_1 \eta_2)^{M-1} u_3^M]$$

satisfies (H1)-(H4), and (H6) with $L_1 = L_2 = K_6 = 0$. Thus if we take M sufficiently large we can apply Proposition 5.3 and Lemma 4.1 to obtain boundedness results if $b_1 = 0$ and decay results if $b_1 \neq 0$.

A preprint by Farr, Fitzgibbon, Morgan, and Waggoner [21] discusses a large family of chemical reactions possessing similar rich Lyapunov structures.

REFERENCES

[1] P. BATES, *Containment of weakly coupled parabolic systems*, Houston J. Math., 11 (1985), pp. 151-158.

[2] P. BATES AND K. BROWN, *Convergence to equilibrium in a reaction-diffusion system*, Nonlinear Anal. Theory Methods Appl., 8 (1984), pp. 227-235.

[3] N. CHUEH, C. CONLEY, AND J. SMOLLER, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373-392.

[4] A. FRIEDMAN AND A. TZAVARES, *A quasilinear parabolic system arising in modelling of catalytic reactors*, J. Differential Equations, 70 (1987), pp. 167-196.

[5] P. GRAY AND S. SCOTT, *Sustained oscillations in a* CSTR, J. Phys. Chem., 89 (1985), p. 22.

[6] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, New York, 1981.

[7] S. HOLLIS, *Globally Bounded Solutions of Reaction-Diffusion Systems*, Ph.D. thesis, North Carolina State University, Raleigh, NC, 1986.

[8] S. HOLLIS, R. MARTIN, AND M. PIERRE, *Global existence and boundedness in reaction-diffusion systems*, SIAM J. Math. Anal., 18 (1987), pp. 744-761.

[9] Y. I. KANEL, *Cauchy's problem for semilinear parabolic equations with balance conditions*, Trans. Differensial'nye Uravneniya, 20 (1984), pp. 1753-1760.

[10] O. LADYZHENSKAJA, V. SOLONNIKOV, AND N. URALCEVA, *Linear and Quasilinear Equations of Parabolic Type*, Amer. Math. Soc. Transl. 23, American Mathematical Society, Providence, RI, 1968.

[11] J. LASRY, *International working paper of the Mathematical Research Center*, CEREMADE, University of Paris-Dauphine, Paris.

[12] J. LIGHTBOURNE AND R. MARTIN, *Relatively continuous nonlinear perturbations of analytic semigroups*, Nonlinear Anal. Theory Methods Appl., 1 (1977), pp. 277-292.

[13] J. MORGAN, *Global existence for semilinear parabolic systems*, SIAM J. Math. Anal., 20 (1989), pp. 1128-1144.

[14] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, Berlin, New York, 1983.

[15] R. REDHEFFER, R. REDLINGER, AND W. WALTER, *A theorem of La Salle-Lyapunov type for parabolic systems*, SIAM J. Math. Anal., 19 (1988), pp. 121-132.

[16] F. ROTHE, *Global Solutions of Reaction-Diffusion Systems*, Lecture Notes in Math. 1072, Springer-Verlag, Berlin, New York, 1984.

[17] M. SINGH, K. KHETARPAL, AND M. SHARAN, *A theoretical model for studying the rate of oxygenation of blood in pulmonary capillaries*, J. Math. Biol., 9 (1980), pp. 305-330.

[18] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, New York, 1983.

[19] J. VASTANO, J. PEARSON, W. HORSTHEMKE, AND H. SWEENEY, *Chemical pattern formation with equal diffusion coefficients*, preprint.

[20] H. WEINBERGER, *Invariant sets for weakly coupled parabolic and elliptic systems*, Rend. Mat., 8 (1975), pp. 295-310.

[21] W. W. FARR, W. E. FITZGIBBON, J. J. MORGAN, AND S. J. WAGGONER, *Boundedness and asymptotic convergence for a class of autocatalytic chemical systems*, preprint, 1990.

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1190-1212, September 1990
© 1990 Society for Industrial and Applied Mathematics
005

ATTRACTORS FOR THE SYSTEM OF SCHRÖDINGER AND KLEIN-GORDON EQUATIONS WITH YUKAWA COUPLING*

PIOTR BILER†

Abstract. A system of two hyperbolic equations describing the interaction of a complex nucleon field with a real meson field is considered in a domain of $\mathbb{R}^n$, $n \leqq 3$. The global finite-dimensional attractor is constructed for slightly damped equations driven by exterior forces. The asymptotic behavior of such attractors is studied for a singular perturbation problem. Some estimates on the decay of solutions of homogeneous systems are also established.

Résumé. On considère un système de deux équations hyperboliques modélisant l'interaction d'un champ complexe de nucléons avec un champ réel de mésons dans un domaine borné de $\mathbb{R}^n$, $n = 1, 2, 3$. On montre l'existence d'un attracteur de dimension finie qui capture toutes les trajectoires du système en présence d'une faible dissipation et de forces extérieures. On étudie aussi un problème de perturbation singulière ainsi que le comportement asymptotique du système homogène.

Key words. global attractor, Klein-Gordon equations, Schrödinger equations, Yukawa coupling, asymptotic behavior of solutions

AMS(MOS) subject classifications. 35B40, 35B99, 35Q20, 81C99

0. Introduction. In this paper we consider the following system of semilinear Schrödinger and Klein-Gordon equations:

$$i\psi_t - A_1\psi + i\varepsilon\psi = -\psi\varphi + F, \tag{0.1}$$

$$\varphi_{tt} + \delta\varphi_t + A_2\varphi = |\psi|^2 + G, \tag{0.2}$$

defined on a bounded domain Ω in $\mathbb{R}^n$, $n = 1, 2, 3$. Here A_1, A_2 are second-order elliptic differential operators, in the simplest case $A_1 = -\Delta$, $A_2 = -\Delta + m^2$. In the conservative case, when $\varepsilon = \delta = 0$, $F = G = 0$, these equations describe the dynamics of a complex nucleon field ψ and a real meson field φ coupled through the Yukawa interaction. Perturbations of this classical semirelativistic system include weak dissipation (of order zero) introduced by the terms $i\varepsilon\psi$, $\delta\varphi_t$ with ε, $\delta > 0$ and driving forces $F = F(x, t)$, $G = G(x, t)$. The system (0.1), (0.2), supplemented with the homogeneous Dirichlet boundary conditions (or space periodicity conditions if $\Omega = [0, 1]^n$), constitutes a model in the field theory with energy dissipation and external excitation in finite volume. The mathematical structure of this system is quite simple (see the system (6.10), (6.11)) but its analysis involves some interesting questions.

The conservative system (and its relativistic generalization) has been studied by many authors. Most of the results concern the Cauchy problem in the whole space $\mathbb{R}^n$, $n = 1, 2, 3$ (cf. [2]). The initial boundary value problems on an open subset of $\mathbb{R}^n$ have been investigated in [8], [9], and [19]. We refer the reader to these papers and to the references therein for a more complete account of the physical significance of the system and the precise statements of the results.

We are interested in the asymptotic behavior of solutions of (0.1), (0.2) when time $t \to \infty$ in the dissipative case, i.e., $\varepsilon > 0$, $\delta > 0$. We expect long-time behavior similar to that for the single nonlinear damped Schrödinger equation or the single nonlinear

* Received by the editors January 12, 1988; accepted for publication (in revised form) November 6, 1989.
† Mathematical Institute, University of Wrocław, pl. Grunwaldzki 2/4, 50-384 Wrocław, Poland. Part of this research was done while the author was at the Laboratoire d'Analyse Numérique, Université de Paris-Sud, 91405 Orsay, France.

dissipative Klein-Gordon equation considered in [11], [1], and [14], i.e., the solutions are asymptotically confined to a finite-dimensional subset of the infinite-dimensional phase space. We will show that the weak dissipation in (0.1), (0.2), which does not suffice to produce regularizing effects on all individual trajectories as for parabolic equations, is sufficient, however, to impose asymptotically finite-dimensional behavior as $t \to \infty$. Namely, there exists a compact set that captures all solutions (Theorem 4.1). This global attractor consists of regular solutions and its fractal dimension is finite (Theorem 8.1).

The outline of this paper is as follows. After introducing the notation, in § 1 we give some remarks on the solvability of the initial boundary value problem following the results in the conservative case due to Hayashi and von Wahl [19] and Fukuda and Tsutsumi [8]. We establish in §§ 2 and 3 some uniform bounds for suitable norms of the solutions and then, in § 4, we apply these estimates to the construction of the global attractor. In §§ 7 and 8 we estimate the dimension of this attractor using the Lyapunov exponents for the flows on compact invariant sets. These techniques presented for parabolic type equations in [7] have been modified and generalized in [14], [11], [12] for damped hyperbolic and Schrödinger (as well as Korteweg-de Vries) equations. In § 9 we study a singular perturbation problem when the damping parameters tend to zero and the corresponding attractors are preserved in some sense when we pass to a singular limit in the Klein-Gordon equation. We also give some estimates on the decay of solutions of the homogeneous dissipative system ($F = G = 0$, § 6) and, improving the results of [19], for the classical Hamiltonian system ($\varepsilon = \delta = 0$, § 5). We will only sketch fragments of standard proofs, whereas we will stress some technical novelties such as intermediate space estimates for hyperbolic equations (3.2), a smoothing effect for globally bounded solutions (Proposition 4.3), a uniqueness result for weak solutions (Proposition 4.1 and Remark 4.1), and H^2 estimates for the conservative system (§ 5).

1. The initial boundary value problem. Here we give assumptions concerning (0.1), (0.2). Let Ω be a bounded domain in $\mathbb{R}^n$, $n = 1, 2, 3$, with smooth boundary. The functions ψ, φ are defined on $\Omega \times \mathbb{R}$ with their values in $\mathbb{C}$, $\mathbb{R}$ respectively, and they satisfy the zero Dirichlet condition on the boundary of Ω for all t. The uniformly elliptic second-order operators A_1, A_2

$$A_k u = -\Sigma_{i,j} \frac{\partial(a_{ij}^{(k)}(x)\partial u/\partial x_j)}{\partial x_i} + a^{(k)}(x)u, \qquad k = 1, 2, \tag{1.1}$$

have smooth coefficients, so they are defined on the spaces W_1, W_2 equal to the complex, respectively, the real subspace $H^2(\Omega) \cap H_0^1(\Omega)$ of the Hilbert space $H = L^2(\Omega)$. We reserve special notation V_1, V_2 for the domains of $A_1^{1/2}$, $A_2^{1/2}$, i.e., $H_0^1(\Omega)$ contained in complex, respectively, real, copy of H. We use the notation $|\cdot|_p$ for $L^p(\Omega)$ norms, $\|\cdot\|_r$ for H^r norms, and $(\cdot, \cdot)$ for the scalar product in $L^2(\Omega)$. The index $p = 2$ is omitted if it does not produce any confusion with the notation for the modulus of a complex number. The lower bound of A_k is denoted by $\lambda_1^{(k)} > 0$

$$\lambda_1^{(k)}|u|^2 \leqq (A_k u, u), \qquad k = 1, 2, \tag{1.2}$$

and λ is the best constant in the inequalities

$$\lambda \|u\|_1^2 \leqq (A_k u, u), \qquad k = 1, 2. \tag{1.3}$$

The spectrum of A_k can be arranged in the nondecreasing sequence $0 < \lambda_1^{(k)} \leqq \lambda_2^{(k)} \leqq \cdots$ of the eigenvalues (counting multiplicities). We will only use the positivity of A_k in

H and the fact that they generate analytic semigroups in H (to apply the interpolation theory); however, the dynamics of the system (0.1), (0.2) is connected rather with unitary groups generated by self-adjoint operators derived from A_2 and A_1.

$C(I, E)$ denotes the space of continuous functions defined on an interval $I \subset \mathbb{R}$ with their values in a Hilbert space E. $C_b(I, E)$ is the subspace of uniformly bounded continuous E-valued functions. In the sequel, different positive constants might be denoted by the same letter C. They do not depend on time. $C(\cdot, \cdots, \cdot)$ are constants depending only on the quantities appearing in parentheses. For the questions concerning frequently used interpolation theory and fractional powers of A_k we refer the reader to Volume 1 of [20].

Two subspaces of $\mathbf{H} = H_{\mathbb{C}} \times H_{\mathbb{R}} \times H_{\mathbb{R}}$ will be of special interest in the functional setting for the initial boundary value problem for (0.1), (0.2), namely, $\mathbf{W} = W_1 \times W_2 \times V_1 \subset \mathbf{V} = V_1 \times V_2 \times H$. The first coordinate in all these spaces is complex, the second and the third are restricted to the real part of H.

Using the Galerkin approximations as in [8], we can prove in a standard way the existence of the weak solutions $\langle \psi, \varphi, \varphi_t \rangle \in L^\infty_{\mathrm{loc}}(\mathbb{R}^+, \mathbf{V})$ of (0.1), (0.2) with $F \in L^\infty_{\mathrm{loc}}(\mathbb{R}, H)$, $F_t \in L^\infty_{\mathrm{loc}}(\mathbb{R}, V_1')$, $G \in L^\infty_{\mathrm{loc}}(\mathbb{R}, H)$ and the initial conditions

$$\langle \psi_0, \varphi_0, \varphi_1 \rangle = \langle \psi(0), \varphi(0), \varphi_t(0) \rangle \in \mathbf{V}. \tag{1.4}$$

This result can be improved up to the continuity of $\langle \psi, \varphi, \varphi_t \rangle$ in time using classical arguments from [20, Chaps. 3.8, 5.12]. The regularity assumption on F, G can be relaxed (cf. [20, Chap. 3.11] for the first equation and [14, Prop. 1.3] for the second equation).

The regularity of the solutions is improved assuming that $F \in L^\infty_{\mathrm{loc}}(\mathbb{R}, V_1)$, $F_t \in L^\infty_{\mathrm{loc}}(\mathbb{R}, H)$, $G \in L^\infty_{\mathrm{loc}}(\mathbb{R}, V_2)$, and

$$\langle \psi_0, \varphi_0, \varphi_1 \rangle \in \mathbf{W}. \tag{1.5}$$

In this situation the integral equations approach from [19] gives the existence of the solution $\langle \psi, \varphi, \varphi_t \rangle \in C(\mathbb{R}^+, \mathbf{W})$ and $\varphi_{tt} \in C(\mathbb{R}^+, H)$. Some uniqueness and continuity properties of the solutions are proved in § 4.

Remark 1.1. Observe that the weak dissipation introduced in the system by the terms $i\varepsilon\psi$, $\delta\varphi_t$ provides no problem with backward continuation of the solutions: they can be defined on the whole real line (see the corresponding results for single equations in [11] and [14]).

2. Energy estimates I. In this section we derive several energy identities and collect them to obtain H^1 estimates uniform in time for the solutions of the problem (0.1), (0.2), and (1.4). The local in time estimates are much easier to obtain and, in fact, have been used to construct and continue the solutions described in the previous section. The following calculations are formal and a rigorous justification of them can be given working with the Galerkin approximations and then passing to the limit.

Although in the study of attractors F and G will be assumed to be time independent, we state the main result of this section in the following form.

PROPOSITION 2.1. *Given F, G such that F, F_t, $G \in C_b(\mathbb{R}^+, H)$, and $\langle \psi_0, \varphi_0, \varphi_1 \rangle \in \mathbf{V}$, there exists a unique solution $\langle \psi, \varphi, \varphi_t \rangle \in C(\mathbb{R}^+, \mathbf{V})$ of the system (0.1), (0.2). Moreover, this solution is uniformly bounded in $\mathbf{V}$: $\langle \psi, \varphi, \varphi_t \rangle \in C_b(\mathbb{R}^+, \mathbf{V})$ with $\limsup_{t \to +\infty} |\langle \psi, \varphi, \varphi_t \rangle|_V \leqq C_V$, where C_V is independent of the initial data.*

Proof. The first energy equation

$$\frac{d|\psi|^2}{dt} + 2\varepsilon|\psi|^2 = 2 \operatorname{Im}(\psi, F) \tag{2.1}$$

is obtained by taking the scalar product of (0.1) with ψ and the imaginary part of the resulting identity. The next one is

$$\frac{d(A_1\psi, \psi)}{dt} = 2 \operatorname{Re}(A_1\psi, \psi_t) = 2 \operatorname{Im}(A_1\psi, \varphi\psi) - 2\varepsilon(A_1\psi, \psi) - 2 \operatorname{Im}(A_1\psi, F), \tag{2.2}$$

and this follows by taking the scalar product of (0.1) with ψ_t. The third equation is reminiscent of the well-known invariant of the classical nonlinear cubic Schrödinger equation

$$\begin{aligned} \frac{d(-|\psi|^2, \varphi)}{dt} &= -(|\psi|^2, \varphi_t) - 2 \operatorname{Re}(\psi_t, \varphi\psi) \\ &= -(|\psi|^2, \varphi_t) - 2 \operatorname{Im}(A_1\psi, \varphi\psi) + 2\varepsilon(|\psi|^2, \varphi) - 2 \operatorname{Im}(F, \varphi\psi). \end{aligned} \tag{2.3}$$

Finally, we have for (0.1)

$$\begin{aligned} \frac{d(\operatorname{Re}(F, \psi))}{dt} &= \operatorname{Re}(F_t, \psi) + \operatorname{Re}(F, \psi_t) \\ &= \operatorname{Re}(F_t, \psi) + \operatorname{Im}(F, \varphi\psi) - \operatorname{Im}(F, A_1\psi) - \varepsilon \operatorname{Re}(F, \psi). \end{aligned} \tag{2.4}$$

Taking the scalar product of (0.2) with φ_t, φ respectively, we get

$$\frac{1}{2}\frac{d|\varphi_t|^2}{dt} + \delta|\varphi_t|^2 + \frac{1}{2}\frac{d(A_2\varphi, \varphi)}{dt} = (|\psi|^2, \varphi_t) + (G, \varphi_t) \tag{2.5}$$

and

$$\frac{d(\varphi_t, \varphi)}{dt} - |\varphi_t|^2 + \delta(\varphi_t, \varphi) + (A_2\varphi, \varphi) = (|\psi|^2, \varphi) + (G, \varphi). \tag{2.6}$$

Estimating the right-hand side of (2.1) using the Cauchy-Schwarz inequality, we get $d|\psi|^2/dt + 2\varepsilon|\psi|^2 \leqq 2|F||\psi|$ so $d|\psi|/dt + \varepsilon|\psi| \leqq |F|$ and the uniform L^2 estimate of ψ follows:

$$|\psi(t)| \leqq \exp(-\varepsilon t)|\psi_0| + (1 - \exp(-\varepsilon t))\varepsilon^{-1} \sup_{t \geqq 0} |F(t)|. \tag{2.7}$$

Now we pass to H^1 estimates. We have as a consequence of (2.2)-(2.4)

$$\begin{aligned} &\frac{d((A_1\psi, \psi) - (|\psi|^2, \varphi) + 2 \operatorname{Re}(F, \psi))}{dt} \\ &\quad + 2\varepsilon(A_1\psi, \psi) - 2\varepsilon(|\psi|^2, \varphi) + 2\varepsilon \operatorname{Re}(F, \psi) = -(|\psi|^2, \varphi_t) + 2 \operatorname{Re}(F_t, \psi). \end{aligned} \tag{2.8}$$

Adding (2.5) and (2.6) multiplied by a small positive number μ ($\mu \leqq \min(\varepsilon, \frac{1}{4}\delta, \frac{1}{2}\lambda_1^{(2)}/\delta)$ will work perfectly in the sequel), after some rearrangements, we get

$$\begin{aligned} &\frac{d(\frac{1}{2}|\varphi_t + \mu\varphi|^2 + \frac{1}{2}(A_2\varphi, \varphi))}{dt} + (\delta - \mu)|\varphi_t + \mu\varphi|^2 + \mu(A_2\varphi, \varphi) \\ &\quad = \mu(\delta - \mu)(\varphi_t + \mu\varphi, \varphi) + (|\psi|^2, \varphi_t + \mu\varphi) + (G, \varphi_t + \mu\varphi). \end{aligned} \tag{2.9}$$

Let us note that this trick with the use of $\varphi_t + \mu\varphi$ instead of φ_t in the energy equation for wave-type equations has been rediscovered by several authors (e.g., [18], [14], [21], [3]) after Rabinowitz, who found it in 1967. It seems to be indispensable in establishing certain exact asymptotics (see [3]-[5] where a further modification is given).

As a consequence of (2.8) and (2.9) we deduce

$$\frac{dZ}{dt}+\mu Z+(2\varepsilon-\mu)(A_1\psi,\psi)+\left(\delta-\frac{3}{2}\mu\right)|\varphi_t+\mu\varphi|^2+\frac{1}{2}\mu(A_2\varphi,\varphi)$$

$$=2\varepsilon(|\psi|^2,\varphi)-2(\varepsilon-\mu)\,\mathrm{Re}\,(F,\psi)+\mu(\delta-\mu)(\varphi_t+\mu\varphi,\varphi) \tag{2.10}$$

$$+(G,\varphi_t+\mu\varphi)+2\,\mathrm{Re}\,(F_t,\psi),$$

where $Z=(A_1\psi,\psi)-(|\psi|^2,\varphi)+2\,\mathrm{Re}\,(F,\psi)+\frac{1}{2}|\varphi_t+\mu\varphi|^2+\frac{1}{2}(A_2\varphi,\varphi)$.

The difficulties connected with the estimation of the nonlinear term $(|\psi|^2,\varphi)$ increase with the dimension. In the sequel we will mainly consider the three-dimensional case because the one- and two-dimensional cases are usually easier to handle.

For $n=1$ we can write

$$|(|\psi|^2,\varphi)|\leqq|\psi|_\infty|\psi||\varphi|\leqq C\|\psi\|_1^{1/2}|\psi|^{3/2}|\varphi|,\quad\text{so}$$

$$2\varepsilon(|\psi|^2,\varphi)\leqq\tfrac{1}{2}\lambda_1^{(1)}\varepsilon|\psi|^2+\tfrac{1}{4}\lambda_1^{(2)}\mu|\varphi|^2+C|\psi|^6 \tag{2.11.1}$$

$$\leqq\tfrac{1}{2}\varepsilon(A_1\psi,\psi)+\tfrac{1}{4}\mu(A_2\varphi,\varphi)+C|\psi|^6,\quad\text{where } C=C(\varepsilon,\mu,\lambda_1^{(1)},\lambda_1^{(2)}).$$

For $n=2$ we have a slightly more complicated estimate as $H^1\not\subset L^\infty$

$$|(|\psi|^2,\varphi)|\leqq|\psi|_{8/3}^2|\varphi|_4\leqq C\|\psi\|_1^{1/2}|\psi|^{3/2}\|\varphi\|_1,\quad\text{so}$$

$$2\varepsilon(|\psi|^2,\varphi)\leqq\tfrac{1}{2}\varepsilon(A_1\psi,\psi)+\tfrac{1}{4}\mu(A_2\varphi,\varphi)+C|\psi|^6 \tag{2.11.2}$$

$$\text{with a constant } C=C(\varepsilon,\mu,\lambda_1^{(1)},\lambda).$$

Finally, for $n=3$ we proceed as in [19] to get

$$|(|\psi|^2,\varphi)|\leqq|\psi|_{12/5}^2|\varphi|_6\leqq C\|\psi\|_1^{1/2}|\psi|^{3/2}\|\varphi\|_1,\quad\text{so}$$

$$2\varepsilon(|\psi|^2,\varphi)\leqq\tfrac{1}{2}\varepsilon(A_1\psi,\psi)+\tfrac{1}{4}\mu(A_2\varphi,\varphi)+C|\psi|^6, \tag{2.11.3}$$

$$\text{where } C=C(\varepsilon,\mu,\lambda_1^{(1)},\lambda).$$

The other terms on the right-hand side of (2.11) are simpler to deal with:

$$\begin{aligned}
&2(\varepsilon-\mu)|\mathrm{Re}\,(F,\psi)|\leqq2(\varepsilon-\mu)|F||\psi|\leqq(\varepsilon-\mu)(A_1\psi,\psi)+(\varepsilon-\mu)|F|^2/\lambda_1^{(1)},\\
&2|\mathrm{Re}\,(F_t,\psi)|\leqq\tfrac{1}{2}\varepsilon(A_1\psi,\psi)+2|F_t|^2/(\varepsilon\lambda_1^{(1)}),\\
&|(G,\varphi_t+\mu\varphi)|\leqq\tfrac{1}{4}\delta|\varphi_t+\mu\varphi|^2+\delta^{-1}|G|^2,\\
&\mu(\delta-\mu)|(\varphi_t+\mu\varphi,\varphi)|\leqq\tfrac{1}{2}(\delta-\mu)|\varphi_t+\mu\varphi|^2+\tfrac{1}{2}(\delta-\mu)\mu^2|\varphi|^2.
\end{aligned} \tag{2.12}$$

These estimates lead to the differential inequality

$$\frac{dZ}{dt}+\mu Z+\left(\frac{1}{4}\delta-\mu\right)|\varphi_t+\mu\varphi|^2+\frac{1}{2}\mu\left(\frac{1}{2}-(\delta-\mu)\mu/\lambda_1^{(2)}\right)(A_2\varphi,\varphi)$$

$$\leqq C(|\psi|^6+|F|^2+|F_t|^2+|G|^2), \tag{2.13}$$

and after the integration to

$$Z(t)\leqq\exp(-\mu t)Z(0)+C, \tag{2.14}$$

where $C=C(\varepsilon,\delta,\mu,\lambda_1^{(1)},\lambda_1^{(2)},\lambda,\sup_{t\geqq0}(|\psi(t)|,|F(t)|,|F_t(t)|,|G(t)|))$. A lower bound on Z

$$\begin{aligned}
4Z&\geqq(A_1\psi,\psi)+2|\varphi_t+\mu\varphi|^2+2(A_2\varphi,\varphi)-C\\
&\geqq c(\|\psi\|_1^2+\|\varphi\|_1^2+|\varphi_t|^2)-C,
\end{aligned} \tag{2.15}$$

with a constant C independent of t and suitably small $c>0$, is another consequence of (2.11.n), (2.12). Therefore the uniform bound on Z (2.14) provides us with H^1 upper bounds on φ and ψ that are independent of the initial data. This finishes the proof of Proposition 2.1.

3. Energy estimates II. Our objective here is to derive uniform in time H^2 estimates of the solutions of (0.1), (0.2) with more regular initial conditions (1.5). We remark that further regularity results (beyond H^2) can be obtained in a similar manner (see [8] for the conservative system).

PROPOSITION 3.1. *Let F, G satisfy the conditions $F, F_t \in C_b(\mathbb{R}^+, H)$, $G \in C_b(\mathbb{R}^+, V_2)$, and $\langle \psi_0, \varphi_0, \varphi_1 \rangle \in \mathbf{W}$. Then the unique solution of the system* (0.1), (0.2) *provided by Proposition* 2.1 *is in $C_b(\mathbb{R}^+, \mathbf{W})$. Moreover, there exists a universal constant C_W independent of the initial data such that* $\limsup_{t\to+\infty} |\langle \psi, \varphi, \varphi_t \rangle|_W \leqq C_W$.

Proof. We proceed formally differentiating (0.1), taking the scalar product of the resulting equation with ψ_t, and separating the imaginary part to get

$$\frac{d|\psi_t|^2}{dt} + 2\varepsilon|\psi_t|^2 = -2 \operatorname{Im}(\psi\varphi_t, \psi_t) + 2 \operatorname{Im}(F_t, \psi_t). \tag{3.1}$$

Evidently, the L^2 estimate of φ_t in Proposition 2.1 is not sufficient now. To get a bound on φ_t in a stronger norm, a rather delicate (at least in the three-dimensional case) procedure is involved. We begin with the energy equation for (0.2) obtained by taking the scalar product of this equation with $B(\varphi_t + \mu\varphi)$, where B is a fractional power of A_2:

$$\begin{aligned} &\frac{1}{2}\frac{d((B(\varphi_t+\mu\varphi), \varphi_t+\mu\varphi)+(A_2\varphi, B\varphi))}{dt} \\ &\qquad +(\delta-\mu)(B(\varphi_t+\mu\varphi), \varphi_t+\mu\varphi)+\mu(A_2\varphi, B\varphi) \\ &\qquad =(B(\varphi_t+\mu\varphi), \mu(\delta-\mu)\varphi+G+|\psi|^2) \\ &\qquad \leqq \frac{3}{4}(\delta-\mu)(B(\varphi_t+\mu\varphi), \varphi_t+\mu\varphi)+\mu^2(\delta-\mu)(B\varphi, \varphi) \\ &\qquad +(\delta-\mu)^{-1}(BG, G)+(\delta-\mu)^{-1}(B(|\psi|^2), |\psi|^2). \end{aligned} \tag{3.2}$$

The second term on the right-hand side is absorbed by $\mu(A_2\varphi, B\varphi)$ for small positive μ, the third is bounded by assumption on G, and the fourth is the most difficult.

First let $B = A_2^{1/2}$ in (3.2). In the three-dimensional case we may use nonlinear interpolation (e.g., Lemma 4 in [19]), which gives the inequality $|A_2^{1/4}|\psi|^2|^2 \leqq C|A_2^{1/2}\psi|^4$. Together with Proposition 2.1 this implies the uniform boundedness of the fourth term in (3.2). The cases $n=1$ and $n=2$ are simpler as it is not necessary to apply this sharp interpolation inequality. Now we integrate the inequality resulting from (3.2), and we obtain uniform boundedness of $|A_2^{3/4}\varphi|$ and $|A_2^{1/4}(\varphi_t+\mu\varphi)|$. In particular, $|A_2^{1/4}\varphi_t|$, $\|\varphi_t\|_{1/2}$, and therefore $|\varphi_t|_3$ are bounded as $H^{1/2} \subset L^3$. Equation (3.1) now gives

$$\frac{d|\psi_t|^2}{dt} + 2\varepsilon|\psi_t|^2 \leqq |\psi|_6|\varphi_t|_3|\psi_t| + |F_t||\psi_t| \leqq \varepsilon|\psi_t|^2 + C(|\psi|_6^2 + |F_t|^2) \tag{3.3}$$

or $d|\psi_t|^2/dt + \varepsilon|\psi_t|^2 \leqq C$ since $H^1 \subset L^6$. The uniform boundedness of $|\psi_t|$ follows after an integration. Returning to the original equation (0.1), we see that $|A_1\psi|$ is also uniformly bounded; note that

$$|\psi\varphi| \leqq |\psi|_4|\varphi|_4 \leqq C\|\psi\|_1^{n/4}|\psi|^{1-n/4}\|\varphi\|_1^{n/4}|\varphi|^{1-n/4} \tag{3.4}$$

hold for $n \leqq 4$. In particular, $|\psi|_\infty$ is uniformly bounded in time since $H^2 \subset L^\infty$ for $n = 1, 2, 3$.

In the second step of exploiting (3.2) we put $B = A_2$ and we estimate $(A_2(|\psi|^2), |\psi|^2)$ by $|A_2^{1/2}(|\psi|^2)|^2 \leqq C|A_2\psi|^4 \leqq C\|\psi\|_2^4$ ((2.18) in [19] gives an even more subtle estimate $|A_2^{1/2}(|\psi|^2)| \leqq C|A_2^{5/8}\psi|^2$). Anyway, (3.2) implies for small $\mu > 0$ the uniform boundedness of $|A_2\varphi|$ and $\|\varphi_t\|_1$. Therefore $\|\varphi\|_2$ and $|\varphi_{tt}|$ are uniformly bounded, the last quantity being estimated directly from (0.2). This accomplishes the proof of Proposition 3.1.

4. Existence and regularity of the global attractor. In the sequel we will consider the autonomous system (0.1), (0.2), i.e., with time-independent forces F, G. However, some results can be generalized to the case of time-periodic excitation F, G. Necessary modifications of proofs that should be done are rather standard so we omit them. They are described in [11] and [14] for Schrödinger and wave equations, respectively.

The results on the unique solvability of the initial boundary value problem for (0.1), (0.2), (1.4), or (1.5) formulated in § 1 may be interpreted as follows. The mapping

$$S(t)(\langle\psi_0, \varphi_0, \varphi_1\rangle) = \langle\psi(t), \varphi(t), \varphi_t(t)\rangle \tag{4.1}$$

is defined for all t and for all the triples $\langle\psi_0, \varphi_0, \varphi_1\rangle \in \mathbf{V}$. $S(t)$ acts on $\mathbf{V}$ and leaves $\mathbf{W}$ invariant. The family of all $S(t)$, $t \in \mathbb{R}$, constitutes a nonlinear group on $\mathbf{V}$ and on $\mathbf{W}$ because the considered system is autonomous. As it concerns continuity properties of the mappings $S(t)$ we prove the following proposition.

PROPOSITION 4.1. *For every t, $S(t)$ is continuous as the mapping in $\mathbf{V}$ equipped with the norm inherited from the space $L^2 \times H^{1/2} \times H^{-1/2}$ containing $\mathbf{V}$. Moreover, $S(t)$ is continuous as the mapping in $\mathbf{W}$ equipped with the usual norm.*

Proof. We consider two solutions $\langle\psi^{(1)}, \varphi^{(1)}\rangle$, $\langle\psi^{(2)}, \varphi^{(2)}\rangle$ of the system (0.1), (0.2) and their differences $\Psi = \psi^{(1)} - \psi^{(2)}$, $\Phi = \varphi^{(1)} - \varphi^{(2)}$ satisfying the following system of equations:

$$i\Psi_t - A_1\Psi + i\varepsilon\Psi = -\varphi^{(2)}\Psi - \psi^{(1)}\Phi, \tag{4.2}$$

$$\Phi_{tt} + \delta\Phi_t + A_2\Phi = \psi^{(1)}\bar{\Psi} + \overline{\psi^{(2)}}\Psi. \tag{4.3}$$

Let us write the counterparts of the energy equations (2.1) and (2.5) for the above system:

$$\begin{aligned} \frac{d|\Psi|^2}{dt} + 2\varepsilon|\Psi|^2 &= 2\,\mathrm{Im}\,(\Psi, \psi^{(1)}\Phi) \leqq C|\Psi||\psi^{(1)}|_6|\Phi|_3 \\ &\leqq C|\Psi|\,\|\Phi\|_{1/2} \leqq C(|\Psi|^2 + (A_2^{1/2}\Phi, \Phi)), \end{aligned} \tag{4.4}$$

$$\begin{aligned} &\frac{1}{2}\frac{d(A_2^{-1/2}\Phi_t, \Phi_t)}{dt} + \delta(A_2^{-1/2}\Phi_t, \Phi_t) + \frac{1}{2}\frac{d(A_2^{1/2}\Phi, \Phi)}{dt} \\ &\quad = (A_2^{-1/4}\Phi_t, A_2^{-1/4}(\psi^{(1)}\bar{\Psi} + \overline{\psi^{(2)}}\Psi)). \end{aligned} \tag{4.5}$$

Now we recall the inequality $\|uv\|_{-1/2} \leqq C|u|_2|v|_6 \leqq C|u|_2\|v\|_1$, which follows easily from the imbeddings $H^{1/2} \subset L^3$, $H^1 \subset L^6$, and $L^2L^3L^6 \subset L^1$. The right-hand side of (4.5) can be estimated by $C|A_2^{-1/4}\Phi_t|(\|\psi^{(1)}\bar{\Psi}\|_{-1/2} + \|\overline{\psi^{(2)}}\Psi\|_{-1/2}) \leqq C|A_2^{-1/4}\Phi_t|(|\Psi|\,\|\psi^{(1)}\|_1 + |\Psi|\,\|\psi^{(2)}\|_1) \leqq C(|A_2^{-1/4}\Phi_t|^2 + |\Psi|^2)$, since $\psi^{(1)}$, $\psi^{(2)}$ are bounded in H^1. After summing up (4.4) and (4.5), the Gronwall inequality applies. This gives the continuous dependence of $\langle\Psi, \Phi, \Phi_t\rangle$ with respect to initial data in the $L^2 \times H^{1/2} \times H^{-1/2}$ norm. In particular, the uniqueness of the solutions with data in $\mathbf{V}$, claimed in § 1, follows.

Remark 4.1. The argument above applies equally to the conservative system for which the uniqueness of the weak solutions was stated as an open problem in [8].

However, we are not able to prove the continuous dependence of ψ on its initial condition in H^1 norm, which is natural in view of solvability and uniqueness results in H^1 (except for the case $n=1$, which is easy due to the imbedding $H^1 \subset L^\infty$). Although we do not know whether $S(t)$ is continuous in the norm of $\mathbf{V}$, the property stated in Proposition 4.1 is sufficient for our purposes.

We turn to the proof of the second part of Proposition 4.1. Let us consider the equations obtained from (4.2), (4.3) by differentiating them with respect to time. Taking the scalar product with Ψ_t, Φ_{tt}, respectively, we get

$$\frac{d|\Psi_t|^2}{dt}+2\varepsilon|\Psi_t|^2=-2\,\mathrm{Im}\,(\Psi_t,\varphi_t^{(2)}\Psi+\varphi^{(2)}\Psi_t+\psi_t^{(1)}\Phi+\psi^{(1)}\Phi_t) \tag{4.6}$$

$$\leqq C(|\Psi_t|^2+\|\Psi\|_1^2+|\Psi_t|^2+\|\Phi\|_2^2+\|\Phi_t\|_1^2),$$

$$\frac{1}{2}\frac{d|\Phi_{tt}|^2}{dt}+\delta|\Phi_{tt}|^2+\frac{1}{2}\frac{d(A_2\Phi_t,\Phi_t)}{dt}$$

$$=(\Phi_{tt},\psi_t^{(1)}\bar{\Psi}+\psi^{(1)}\overline{\Psi_t}+\overline{\psi_t^{(2)}}\Psi+\overline{\psi^{(2)}}\Psi_t) \tag{4.7}$$

$$\leqq C|\Phi_{tt}|(\|\Psi\|_2+|\Psi_t|)\leqq C(|\Phi_{tt}|^2+\|\Psi\|_2^2+|\Psi_t|^2).$$

Observe that the expressions $(|\Psi_t|^2+|\Phi_{tt}|^2+\|\Phi_t\|_1^2+|\Psi|^2)^{1/2}$ and $(\|\Psi\|_2^2+\|\Phi_t\|_1^2+\|\Phi\|_2^2)^{1/2}$ are equivalent as norms on $\mathbf{W}$. Now a standard argument involving the Gronwall inequality applied to a linear combination of (4.4), (4.6), (4.7) proves the continuity of $S(t)$ in $\mathbf{W}$. The family of mappings $S(t)$, $t\in\mathbb{R}$, acts in $\mathbf{W}$ as a group of homeomorphisms. The extension of $S(t)$ for negative t and the fact that each $S(t)$ is a homeomorphism follow from the backward solvability (see Remark 1.1).

Remark 4.2. To prove another continuity property of $S(t)$, let us consider the higher-order analogues of (4.4), (4.5)

$$\frac{d(B_1\Psi,\Psi)}{dt}+2\varepsilon(B_1\Psi,\Psi)=2\,\mathrm{Im}\,(B_1^{1/2}\Psi,B_1^{1/2}(\varphi^{(2)}\Psi+\psi^{(1)}\Phi)), \tag{4.8}$$

$$\frac{1}{2}\frac{d(B_2\Phi_t,\Phi_t)}{dt}+\delta(B_2\Phi_t,\Phi_t)+\frac{1}{2}\frac{d(A_2\Phi,B_2\Phi)}{dt} \tag{4.9}$$

$$=(B_2^{1/2}\Phi_t,B_2^{1/2}(\psi^{(1)}\bar{\Psi}+\overline{\psi^{(2)}}\Psi)),$$

where $B_k=A_k^s$ with a positive exponent s, $k=1,2$. Using the fact that for $s>n/2$, H^s is a Banach algebra with pointwise multiplication and the Gronwall inequality, we immediately obtain the continuous dependence with respect to initial data in the $H^s\times H^{s+1}\times H^s$ norm for $s>n/2$.

The results stated in Propositions 2.1 and 3.1 may be interpreted as the existence of bounded absorbing sets B_V, B_W for the semigroup $(S(t)\colon t\geqq 0)$ in $\mathbf{V}$, respectively, $\mathbf{W}$. Namely, if we define B_V and B_W as the balls in $\mathbf{V}$, $\mathbf{W}$ of radii C_V+1, C_W+1 respectively, centered at the origin, then the following fact holds.

PROPOSITION 4.2. *For every bounded set of initial data B in $\mathbf{V}$ (respectively, $\mathbf{W}$) there exists time $T=T(B)$ such that $S(t)(B)\subset B_V$ (B_W, respectively) for all $t\geqq T$.*

Our next step is the construction of the global attractor achieved in the following theorem.

THEOREM 4.1. *Under the hypotheses of Proposition* 2.1 *with F, G independent of t, the set*

$$\mathscr{A}=\bigcap_{s\geqq 0}\overline{\bigcup_{t\geqq s}S(t)(B_V)}^{wV}, \tag{4.10}$$

where the closures are taken with respect to the weak topology in $\mathbf{V}$, *is the global attractor for the system* (0.1), (0.2). $\mathscr{A}$ *is nonvoid, connected, and compact in the norm topology of* $\mathbf{V}$, $S(t)$ *invariant, and attracts bounded sets in* $\mathbf{V}$:

$$\begin{aligned} &\varnothing \neq \mathscr{A} \subset B_W, \qquad S(t)(\mathscr{A}) = \mathscr{A} \quad \text{for all } t \in \mathbb{R}, \\ &\lim_{t \to +\infty} \operatorname{dist}(S(t)(B), \mathscr{A}) = 0 \quad \text{for each bounded set } B \text{ in } \mathbf{V} \end{aligned} \tag{4.11}$$

(*the distance is measured with respect to any metric compatible with the weak topology restricted to bounded subsets of* $\mathbf{V}$).

The formulation of Theorem 4.1 calls for some comments.

Remark 4.3. Of course, we are tempted to define an attractor $\mathscr{A}$ as the ω-limit set of a bounded absorbing set, i.e., $\mathscr{A} = \bigcap_{s \geq 0} \overline{\bigcup_{t \geq s} S(t)(B)}$. However, if we would like to construct the global attractor for the solutions with initial conditions in the space $\mathbf{V}$, hence to take $B = B_V$, it is not immediately clear in what topology this construction will work. The natural choice of the norm topology in $\mathbf{V}$ is suitable only in the one-dimensional case when the continuity of $S(t)$ in this topology is available. In the two- and three-dimensional cases we should use weak topologies for two different technical reasons: the lack (of the proof) of continuity of $S(t)$ in $\mathbf{V}$, and the lack of an asymptotic (when $t \to +\infty$) smoothness property of the group $S(t)$, similar to that described in Lemma 3.1 of [14]. This last property may fail because we can prove merely V_1' regularity of ψ_t when $\psi \in C(\mathbb{R}^+, V_1)$ (however, see Proposition 4.3).

The proof of Theorem 4.1 is based on the following technical Lemmas 4.1 and 4.2, and Proposition 4.3.

LEMMA 4.1. *For each* t, $S(t)$ *is continuous with respect to the weak topology of* $\mathbf{V}$.

The proof is essentially the same as that of Proposition 2.2 in [11] (see also (2.7), (2.8) in [11] and (4.4) in this section).

LEMMA 4.2 [14, Prop. 3.1]. *Let* $(\mathbf{V}, d)$ *be a metric space. If the semigroup* $S(t)$ *possesses a bounded absorbing set* B_V *in* $\mathbf{V}$ *and for every bounded set* B *in* $\mathbf{V}$ *there exists a compact set* $K \subset \mathbf{V}$ *such that* $\lim_{t \to +\infty} \sup_B \operatorname{dist}(S(t)(\langle \psi, \varphi, \varphi_t \rangle), K) = 0$, *then* $\mathscr{A} = \bigcap_{s \geq 0} \bigcup_{t \geq s} S(t)(B_V)^d$ *is the global attractor in* $(\mathbf{V}, d)$.

PROPOSITION 4.3. *For* $F \in C_b(\mathbb{R}, W_1)$, $F_t \in C_b(\mathbb{R}, H)$, *and* G, $G_t \in C_b(\mathbb{R}, H)$ *any solution* $\langle \psi, \varphi, \varphi_t \rangle$ *of* (0.1), (0.2) *that is defined for all real* t *and belongs to* $C_b(\mathbb{R}, \mathbf{V})$ *is in fact in* $C_b(\mathbb{R}, \mathbf{W})$. *Its norm in* $C_b(\mathbb{R}, \mathbf{W})$ *is estimated by a quantity depending continuously on its norm in* $C_b(\mathbb{R}, \mathbf{V})$ *and the norms of* F, G.

The proof of this proposition is postponed to the end of the proof of Theorem 4.1.

Proof of Theorem 4.1. B_V is a bounded absorbing set in $\mathbf{V}$, so we have information on all trajectories starting from the initial conditions in $\mathbf{V}$, since they eventually enter in B_V. The weak topology in $\mathbf{V}$ restricted to bounded subsets in $\mathbf{V}$ is metrizable. We take as the metric d in Lemma 4.2 any metric compatible with this topology and we put $K = \mathscr{A}$, which is compact in d. Since the weak convergence in B_V and the continuity properties of (0.2) established in Proposition 4.1 imply, e.g., the convergence in the norm of $H^s \times H^s \times H^{s-1}$ for all $s < 1$, we may interpret the statement in Theorem 4.1 as the convergence in each space mentioned above.

To prove the compactness of $\mathscr{A}$ in the norm topology of $\mathbf{V}$ we use Proposition 4.3, which is a result of asymptotic smoothness of the globally (forward and backward in time) bounded trajectories. Indeed, $\mathscr{A}$ contains, together with each element $\langle \psi, \varphi, \varphi_t \rangle$ whole, its trajectory: forward and backward. Proposition 4.3 implies that such a trajectory of $S(t)$ is bounded not only in $\mathbf{V}$, as $\mathscr{A}$ is bounded in $\mathbf{V}$, but also in $\mathbf{W}$. Therefore $\mathscr{A}$ is bounded in $\mathbf{W}$ and thus compact in $\mathbf{V}$. The argument above is described in full detail in the proof of Theorem 3.1 of [14]. The attractor $\mathscr{A}$ is connected as the

intersection of a decreasing family of compact connected sets (see, e.g., [11, Lemma 2.1]).

The property described in Proposition 4.3 is somewhat analogous to (3.2) in [14] for a single nonlinear wave equation. A similar fact was used in [18] to improve the results of [1]. Later it was discussed in detail in [14]. It may be interpreted as an extra regularity of solutions bounded on $\mathbb{R}$; it is plausible as the trajectories passing at $t=0$ through a point in $\mathbf{V}\backslash\mathbf{W}$ seem to be "large at $t=-\infty$" (unbounded in $\mathbf{V}$ when $t\to-\infty$) due to the damping. This property serves as a substitute of smoothing properties known for parabolic-type equations and systems in the case of noncompact (semi)groups $S(t)$. It is a very interesting problem whether it holds for other damped equations that are, in the undamped case, Hamiltonian systems such as (genuinely) nonlinear Schrödinger equations or Korteweg-de Vries equations. The miracle that Proposition 4.3 holds in our situation is perhaps connected with the bilinear structure of nonlinearities.

Proof of Proposition 4.3. We consider (0.1) as the linear nonhomogeneous Schrödinger equation in the space V_1

$$\psi_t+\varepsilon\psi+i(A_1-\varphi)\psi=-iF, \tag{4.12}$$

with time-dependent potential $\varphi\in C_b(\mathbb{R}, V_2)$. If we remove the damping term $\varepsilon\psi$ from (4.12) by considering the equation for the new variable $\tilde{\psi}(t)=\exp(\varepsilon t)\psi(t)$, then we can verify the assumptions (8.1), (8.2) in [20, Chap. 3] that guarantee the existence of solutions to such Schrödinger-type equations. Namely, $|(\varphi_t, u, v)|\leqq|\varphi_t||u|_4|v|_4\leqq C|\varphi_t|\,\|u\|_1^{3/4}|u|^{1/4}\|v\|_1^{3/4}|v|^{1/4}$ holds for all $u, v\in H^1$ (cf. (3.4)) and the coercivity condition is satisfied due to (1.3) and (2.11.n). Hence there exists the family of the linear operators $P(t, s)$ solving (4.12) with $F\equiv 0$ and the initial condition set at time s. It is quite easy to prove, using the estimates following (10.12) in [20, Chap. 3], that the operators $\exp(c(t-s))P(t, s)$ are uniformly bounded for $t\geqq s$ for some $c>0$, e.g., $c=\varepsilon/2$. The unique (!) solution of (4.12) in $C_b(\mathbb{R}, V_1)$ is given by the formula

$$\psi(t)=-\int_{-\infty}^{t} iP(t, s)F(s)\,ds=-\int_0^{\infty} iP(t, t-s)F(t-s)\,ds; \tag{4.13}$$

this is the solution with zero condition at $-\infty$. Due to subexponential behavior of $P(t, s)$ this integral is well defined in V_i (see a similar argument in [14, Prop. 1.2]).

Now observe that $\psi_t+iA_1\psi=i(\varphi\psi-F)-\varepsilon\psi$; hence the uniform boundedness of ψ_t in L^2 would imply that of $A_1\psi$. Differentiating (4.13) formally, we get $\psi_t(t)=-\int_0^{\infty} i(P(t, t-s)\,dF(t-s)/dt+(dP(t, t-s)/dt)F(t-s))\,ds$. The first integral term is in $C_b(\mathbb{R}, H)$ due to the regularity assumption on F_t. The second is also in $C_b(\mathbb{R}, H)$, which follows from the assumption on F and from a regularity result for time-dependent Schrödinger equations [20, Chap. 5, Thm. 12.1]. Finally, $\psi_t\in C_b(\mathbb{R}, H)$, so ψ belongs to $C_b(\mathbb{R}, W_1)$.

Returning to (0.2) we observe that $\xi=\varphi_t$ satisfies the equation $\xi_{tt}+\delta\xi_t+A_2\xi=\psi_t\bar{\psi}+\overline{\psi_t}\psi+G_t$ with the right-hand side belonging to $C_b(\mathbb{R}, H)$ (as $H^2\subset L^\infty$). Therefore $\langle\xi, \xi_t\rangle$ belongs to $C_b(\mathbb{R}, V_2\times H)$; hence $\varphi_{tt}\in C_b(\mathbb{R}, H)$ and $\varphi\in C_b(\mathbb{R}, W_2)$. We refer the reader to [14, Prop. 1.2, (5.10)] to compare this with a similar argument for a single equation.

Remark 4.4. The system (0.1), (0.2) has intermediate properties between those of two important physical models: the system of two Klein-Gordon equations, and the system of two Schrödinger equations. Our Theorem 4.1 also has an intermediate character: the result is placed between the construction of attractors for hyperbolic second-order equations (see [14]) in norm topologies of the phase space and for nonlinear Schrödinger equations (see [11]), now available only in weak topologies.

Starting with the weak topology definition (4.10) of the global attractor, Proposition 4.3 gives us the compactness of this attractor in much stronger topology.

Remark 4.5. We may also begin the construction of an attractor by considering $\mathscr{A}^0 = \bigcap_{s\geqq 0} \bigcup_{t\geqq s} S(t)(B_W)^{wW}$, where the closures are taken with respect to the weak topology of **W**. The condition in Lemma 4.1 is fulfilled. Also, Proposition 4.3 shows that this "small" attractor constructed for solutions emerging from the data in **W** coincides with the global attractor $\mathscr{A}$ constructed for the larger class of initial data, in fact, the largest one in which existence and uniqueness are assured. The usual construction of the global attractor in the norm topology of **W** raises an important and difficult question of compactness in **W** of $\bigcap_{s\geqq 0} \overline{\bigcup_{t\geqq s} S(t)(B_W)^W}$.

5. Remarks on the conservative system. The estimates obtained in §§ 2 and 3 allow us to give another (and simpler) proof of local in time estimates of strong solutions of the conservative system with $\varepsilon = \delta = 0$, $F = G = 0$ constructed in [19]. Hayashi and von Wahl do not write them explicitly, but a careful analysis of their proof gives only the bounds such as $\exp(Ct^2)$ in [19, (3.12)] and $\exp(\exp(Ct^2))$ for the norms $\|\psi\|_2$, $\|\varphi\|_2$ in [19, (3.13)-(3.15)], while ours are polynomial in t. Moreover, their method uses in an essential way, besides standard Gronwall lemma arguments, a nontrivial inequality of Brézis and Gallouët [6].

PROPOSITION 5.1. *For every solution of the conservative system* $|\langle\psi, \varphi, \varphi_t\rangle|_V$ *is bounded and* $|\langle\psi, \varphi, \varphi_t\rangle(t)|_W = \mathcal{O}(t^2)$ *as* $t\to\infty$.

Proof. We rewrite (2.1) as

$$\frac{d|\psi|^2}{dt} = 0, \tag{5.1}$$

which immediately gives $|\psi|^2 \leqq C$. Equations (2.8) and (2.5) are now simplified to

$$\frac{d((A_1\psi, \psi) - (|\psi|^2, \varphi))}{dt} = -(|\psi|^2, \varphi_t), \tag{5.2}$$

$$\frac{1}{2}\frac{d(|\varphi_t|^2 + (A_2\varphi, \varphi))}{dt} = (|\psi|^2, \varphi_t), \tag{5.3}$$

so

$$2(A_1\psi, \psi) + |\varphi_t|^2 + (A_2\varphi, \varphi) = 2(|\psi|^2, \varphi) + C \tag{5.4}$$

holds. The nonlinear term is estimated by

$$\begin{aligned} |(|\psi|^2, \varphi)| &\leqq |\psi|^2_{12/5}|\varphi|_6 \leqq C\|\psi\|_1^{n/6}|\psi|^{2-n/6}\|\varphi\|_1 \\ &\leqq c_1\|\psi\|_1^2 + c_2\|\varphi\|_1^2 + C|\psi|^{2(12-n)/(6-n)}, \end{aligned} \tag{5.5}$$

with arbitrarily small positive constants c_1, c_2, and a large constant C. The uniform boundedness of $\langle\psi, \varphi, \varphi_t\rangle$ in **V** is a direct consequence of (5.1), (5.4), (5.5):

$$(A_1\psi, \psi) + (A_2\varphi, \varphi) + |\varphi_t|^2 \leqq C. \tag{5.6}$$

The analogue of (3.2) with $\mu = 0$ is

$$\frac{d((B\varphi_t, \varphi_t) + (A_2\varphi, B\varphi))}{dt} = 2(|\psi|^2, B\varphi_t). \tag{5.7}$$

Putting $B = A_2^{1/2}$ and estimating the right-hand side of (5.7) with Lemma 4 in [19], by $2|B^{1/2}(|\psi|^2)|(B\varphi_t, \varphi_t)^{1/2} \leqq C|A_2^{1/2}\psi|^2(B\varphi_t, \varphi_t)^{1/2} \leqq C\|\psi\|_1^2(B\varphi_t, \varphi_t)^{1/2}$ we get from (5.5)

the inequality $d((A_2^{1/2}\varphi_t, \varphi_t)+(A_2\varphi, A_2^{1/2}\varphi))/dt \leqq C(A_2^{1/2}\varphi_t, \varphi_t)^{1/2}$. After an integration this reads

$$\|\varphi_t(t)\|_{1/2}^2+\|\varphi(t)\|_{3/2}^2=\mathcal{O}(t^2) \quad \text{for } |t|\to\infty. \tag{5.8}$$

Now (3.3) implies $d|\psi_t|^2/dt=-2\,\mathrm{Im}\,(\psi\varphi_t, \psi_t)\leqq|\psi_t||\psi|_6|\varphi_t|_3\leqq C|\psi_t|\,\|\psi\|_1\|\varphi_t\|_{1/2}\leqq Ct|\psi_t|$; hence

$$|\psi_t(t)|_2=\mathcal{O}(t^2). \tag{5.9}$$

It follows from (0.1) that $|A_1\psi|\leqq|\psi_t|+|\psi\varphi|\leqq|\psi_t|+C\|\psi\|_1\|\varphi\|_1$ so

$$\|\psi(t)\|_2=\mathcal{O}(t^2). \tag{5.10}$$

Finally, we return to (5.7), this time with $B=A_2$, to obtain

$$\begin{aligned} d((A_2\varphi_t, \varphi_t)+|A_2\varphi|^2)/dt &\leqq 2(A_2\varphi_t, \varphi_t)^{1/2}|A_2^{1/2}(|\psi|^2)| \\ &\leqq C(A_2\varphi_t, \varphi_t)^{1/2}|A_2^{5/8}\psi|^2 \\ &\leqq C(A_2\varphi_t, \varphi_t)^{1/2}|A_1\psi|^{1/2}|A_1^{1/2}\psi|^{3/2} \end{aligned}$$

using nonlinear interpolation as in (2.18) of [19]. An immediate consequence of this and (0.2) reads

$$\|\varphi_t\|_1+\|\varphi\|_2+|\varphi_{tt}|_2=\mathcal{O}(t^2). \tag{5.11}$$

We indicate that in the case of $n=1$ and $n=2$ the bounds corresponding to (5.9)-(5.11) are of the form

$$|\psi_t|_2+\|\psi\|_2+|\varphi_{tt}|_2+\|\varphi_t\|_1+\|\varphi\|_2\begin{cases}=\mathcal{O}(t) & \text{if } n=1,\\ =\mathcal{O}(t^{1+a}) & \text{for any } a>0 \quad \text{if } n=2.\end{cases} \tag{5.12}$$

Remark 5.1. A natural question arises in this context. Are these norms bounded in time or not? The similar question for the nonlinear Schrödinger equation on bounded domains seems to be far more difficult. In the two-dimensional case the iterated exponential bound given by Brézis and Gallouët in [6] has not yet been improved.

6. Exponential decay of solutions of the damped homogeneous system. The aim of this section is twofold: the results on the optimal decay of solutions of (0.1), (0.2) with $F=G=0$ are interesting on their own, but they also give insight into the behavior of the linearized system (i.e., the equations in variations for (0.1), (0.2)) studied in § 7. We may expect that for the damped homogeneous system the nonlinear terms would be negligible for large time and the asymptotics of solutions would be close to that of the system (0.1), (0.2) linearized about $\psi=\varphi=0$, that is,

$$|\psi(t)|_2=\mathcal{O}(\exp(-\varepsilon t)) \tag{6.1}$$

and $|\varphi(t)|_2=\mathcal{O}(\exp(-\frac{1}{2}\delta t))$ for $t\to+\infty$. However, a correction to this statement should be made: if the right-hand term $|\psi(t)|^2$ in (0.2) is of order $\exp(-2\varepsilon t)$, then, in general, the decay of φ is not better than

$$|\varphi(t)|_2=\mathcal{O}(\exp(-\rho t)) \quad \text{with } \rho=\min(2\varepsilon, \tfrac{1}{2}\delta) \quad (\text{if } 2\varepsilon\neq\tfrac{1}{2}\delta), \tag{6.1'}$$

as is clearly implied by the Duhamel formula for (0.2). Moreover, we should suppose that δ is small: $\delta^2<4\lambda_1^{(2)}$, to exclude the overdamping phenomenon (cf. [3, Thm. (ii), (iii), Remarks 3, 4]). We will prove stronger H^1 versions of these decay estimates.

PROPOSITION 6.1. *For the homogeneous system* (0.1), (0.2) ($F=G=0$) *with* $\delta^2<4\lambda_1^{(2)}$ $\|\psi(t)\|_1=\mathcal{O}(\exp(-\varepsilon t))$ *and* $\|\varphi(t)\|_1=\mathcal{O}(\exp(-\min(2\varepsilon, \frac{1}{2}\delta)t))$ *when* $t\to+\infty$, *except for the resonance case* $2\varepsilon=\frac{1}{2}\delta$ *where* $\|\varphi(t)\|_1=\mathcal{O}(t\exp(-\frac{1}{2}\delta t))$.

Proof. Equation (6.1) is simply a consequence of (2.1): since $F=0$ we have $d|\psi|^2/dt+2\varepsilon|\psi|^2=0$. Summing (2.8) and (2.1) multiplied by a positive parameter ν, we obtain

$$\frac{d((A_1\psi,\psi)-(|\psi|^2,\varphi)+\nu|\psi|^2)}{dt}+2\varepsilon((A_1\psi,\psi)-(|\psi|^2,\varphi)+\nu|\psi|^2)=-(|\psi|^2,\varphi_t). \tag{6.2}$$

Denoting the quantity in parentheses by Y, we easily see that for sufficiently large ν, Y is equivalent to $\|\psi\|_1^2$. There exists a constant $C>0$ such that

$$C^{-1}\|\psi\|_1^2\leqq Y\leqq C\|\psi\|_1^2. \tag{6.3}$$

This follows from the boundedness of φ established in Proposition 2.1 and from the inequality

$$|(|\psi|^2,\varphi)|\leqq|\varphi||\psi|_4^2\leqq C\|\psi\|_1^{n/2}|\psi|^{2-n/2}\leqq a\|\psi\|_1^2+C(a)|\psi|^2, \tag{6.4}$$

with a suitably small $a>0$. Because $|\varphi_t|$ is also bounded we have $|(|\psi|^2,\varphi_t)|\leqq|\varphi_t||\varphi|_4^2\leqq CY^{n/4}|\psi|^{2-n/2}$ for $n=1,2,3$. Now we derive from (6.1), (6.2) the inequality $dY/dt+2\varepsilon Y\leqq CY^{n/4}\exp((\frac{1}{2}n-2)\varepsilon t)$ or $d(\exp(2\varepsilon t)Y)/dt\leqq C(\exp(2\varepsilon t)Y)^{n/4}$. After the integration this gives

$$Y(t)=\mathcal{O}(t^{4/(4-n)}\exp(-2\varepsilon t)). \tag{6.5}$$

Now we recall (2.9) with $G=0$ and $\mu=\frac{1}{2}\delta$ in a slightly transformed version

$$\begin{aligned}&\frac{d(|\varphi_t+\frac{1}{2}\delta\varphi|^2+((A_2-\frac{1}{4}\delta^2)\varphi,\varphi))}{dt}+\delta\left(\left|\varphi_t+\frac{1}{2}\delta\varphi\right|^2+\left(\left(A_2-\frac{1}{4}\delta^2\right)\varphi,\varphi\right)\right)\\&\quad=\left(|\psi|^2,\varphi_t+\frac{1}{2}\delta\varphi\right).\end{aligned} \tag{6.6}$$

Let us denote the quantity in parentheses by J. Since $\delta^2<4\lambda_1^{(2)}$, $J^{1/2}$ is equivalent to the norm in $V_2\times H$

$$C^{-1}(\|\varphi\|_1^2+|\varphi_t|^2)\leqq J\leqq C(\|\varphi\|_1^2+|\varphi_t|^2) \tag{6.7}$$

for some $C>0$. From (6.6), the Cauchy-Schwarz inequality, and (6.5) we infer $dJ/dt+\delta J\leqq CJ^{1/2}t^{4/(4-n)}\exp(-2\varepsilon t)$, or in a more convenient form

$$\frac{dJ^{1/2}}{dt}+\frac{1}{2}\delta J^{1/2}=\mathcal{O}(t^{4/(4-n)}\exp(-2\varepsilon t)). \tag{6.8}$$

The integration of (6.8) leads to

$$J(t)=\begin{cases}\mathcal{O}(\exp(-\delta t)) & \text{if } \delta<4\varepsilon,\\ \mathcal{O}(t^{2(8-n)/(4-n)}\exp(-\delta t)) & \text{if } \delta=4\varepsilon,\\ \mathcal{O}(t^{8/(4-n)}\exp(-4\varepsilon t)) & \text{if } \delta>4\varepsilon.\end{cases} \tag{6.9}$$

Inserting this exponential estimate of φ_t into (6.2), we can improve (6.5) up to $Y(t)=\mathcal{O}(\exp(-2\varepsilon t))$. This bootstrap argument then improves (6.9) up to $J(t)=\mathcal{O}(t^2\exp(-\delta t))$ if $\delta=4\varepsilon$ and $J(t)=\mathcal{O}(\exp(-4\varepsilon t))$ if $\delta>4\varepsilon$.

In the remainder of this section we consider the following homogeneous system of Schrödinger and Klein-Gordon equations:

$$i\psi_t-A_1\psi+i(\eta A_1+\varepsilon)\psi=-\psi\varphi, \tag{6.10}$$

$$\varphi_{tt}+\delta\varphi_t+A_2\varphi=|\psi|^2, \tag{6.11}$$

with modified strong damping term $i(\eta A_1+\varepsilon)\psi$ in the first equation. This system resembles the Ginzburg-Landau equation studied in [13]. We will show that this kind of damping mechanism causes more regular decay of ψ compared to that in Proposition 6.1. From a purely technical point of view, it may be interesting to compare some estimates given here with those in [13], where only a two-dimensional problem is considered but the nonlinearities are stronger. Our main result is Proposition 6.2.

PROPOSITION 6.2. *Let* $\langle\psi, \varphi\rangle$ *be any nonzero solution of the system* (6.10), (6.11). *Then there exists the limit* $\Lambda_\infty=\lim_{t\to+\infty}(A_1\psi(t), \psi(t))/|\psi(t)|^2$ *and* Λ_∞ *belongs to the spectrum of* A_1. *Moreover,* $(A_1\psi(t), \psi(t))$ *is equivalent to a constant multiple of the exponential function* $\exp(-2(\eta\Lambda_\infty+\varepsilon)t)$ *when* $t\to+\infty$.

Remark 6.1. Observe that even with a stronger damping term in (6.11), say $(\nu A_2+\delta)\varphi_t$, there is no hope for obtaining regular exponential decay (similar to that for ψ) of φ. An explanation of this can be deduced from [4] and [5].

Proof of Proposition 6.2. We will not discuss here the questions of existence of solutions that are quite standard, having in mind, e.g., [13]. The counterparts of the energy identities (2.1)-(2.3) are

$$\frac{d|\psi|^2}{dt}+2\eta(A_1\psi, \psi)+2\varepsilon|\psi|^2=0, \tag{6.12}$$

$$\frac{d(A_1\psi, \psi)}{dt}+2\eta|A_1\psi|^2+2\varepsilon(A_1\psi, \psi)=2\,\mathrm{Im}\,(A_1\psi, \varphi\psi), \tag{6.13}$$

$$\frac{d(-|\psi|^2, \varphi)}{dt}=-(|\psi|^2, \varphi_t)-2\,\mathrm{Im}\,(A_1\psi, \varphi\psi)+2\eta\,\mathrm{Re}\,(A_1\psi, \varphi\psi)+2\varepsilon(|\psi|^2, \varphi). \tag{6.14}$$

Clearly, (6.12) implies

$$\begin{aligned}&|\psi(t)|^2+2\eta\int_0^t(A_1\psi(s), \psi(s))\,ds\leqq C,\\ &|\psi(t)|^2=\mathcal{O}(\exp(-2(\eta\lambda_1^{(1)}+\varepsilon)t)).\end{aligned} \tag{6.15}$$

Multiplying (6.12) by a positive number ν and summing up with (6.13), (6.14), we arrive at

$$\begin{aligned}&\frac{d((A_1\psi, \psi)+\nu|\psi|^2-(|\psi|^2, \varphi))}{dt}\\ &\quad+2(\eta\nu+\varepsilon)(A_1\psi, \psi)+2\eta|A_1\psi|^2+2\varepsilon\nu|\psi|^2-2\varepsilon(|\psi|^2, \varphi)\\ &\quad=-(|\psi|^2, \varphi_t)-2\eta\,\mathrm{Re}\,(A_1\psi, \varphi\psi).\end{aligned} \tag{6.16}$$

Recalling (2.5), we have $d(|\varphi_t|^2+(A_2\varphi, \varphi))/dt\leqq C|\varphi_t||\psi|_4^2\leqq C|\varphi_t|\,\|\psi\|_1^{n/2}|\psi|^{2-n/2}$. The Cauchy-Schwarz inequality combined with (6.15) implies the inequality $dJ/dt(t)\leqq CJ^{1/2}(t)I(t)$, where $J(t)=|\varphi_t|^2+(A_2\varphi, \varphi)$ and $I(t)$ is integrable over $\mathbb{R}^+$. This produces a uniform bound on J; hence

$$\|\varphi\|_1+|\varphi_t|\leqq C \quad \text{with } C \text{ independent of } t. \tag{6.17}$$

Now we proceed as in § 2, so we recall (2.9), which gives

$$\begin{aligned}&\frac{1}{2}\frac{d(|\varphi_t+\mu\varphi|^2+(A^2\varphi, \varphi))}{dt}+(\delta-\mu)|\varphi_t+\mu\varphi|^2+\mu(A_2\varphi, \varphi)\\ &\quad=\mu(\delta-\mu)(\varphi_t+\mu\varphi, \varphi)+(|\psi|^2, \varphi_t)+\mu(|\psi|^2, \varphi),\end{aligned} \tag{6.18}$$

and sum it up with (6.16). The new essential difficulty is the term $2\eta|\mathrm{Re}\,(A_1\psi, \varphi\psi)| \leqq \eta|A_1\psi|^2 + \eta|\varphi\psi|^2$, which should be absorbed by $2\eta\nu(A_1\psi, \psi)$, $2\eta|A_1\psi|^2$, $\frac{1}{4}\mu(A_2\varphi, \varphi)$. The three-dimensional case is again the most difficult:

$$\begin{aligned}\eta|\psi\varphi|^2 &\leqq \eta|\varphi|_3^2|\psi|_6^2 \leqq C\|\psi\|_2^{2/3}|\psi|^{2-n/3}\|\varphi\|_1^{n/3}|\varphi|^{2-n/3} \\ &\leqq c\|\varphi\|_1^2|\varphi|^{6/n(2-n/3)} + C\|\psi\|_2^{2n/(6-n)}|\psi|^{6(2-n/3)/(6-n)},\end{aligned} \tag{6.19}$$

with suitably small $c>0$ and some $C>0$. The first term can be made less than $\frac{1}{4}\mu(A_2\varphi, \varphi)$ and the second is less than $\eta(A_1\psi, \psi)$ if we take t sufficiently large. This observation follows from (6.15) and the inequality $2n/(6-n) \leqq 2$. Repeating the arguments in the proof of Proposition 6.1, (6.19) and (6.16) combined with (6.18) yield the exponential decay of $\|\psi\|_1$, $\|\varphi\|_1$, and $|\varphi_t|$ with the exponent $-\frac{1}{2}\mu$.

This preliminary estimate enables us to conclude the proof of Proposition 6.2 in much the same way as in, e.g., [4] and [5]. Namely, we consider the ratio $\Lambda(t) = (A_1\psi(t), \psi(t))/|\psi(t)|^2 =: N(t)/D(t)$, which satisfies the differential equation $D\,d\Lambda/dt + 2\eta|(A_1-\Lambda)\psi|^2 = 2\,\mathrm{Im}\,((A_1-\Lambda)\psi, \varphi\psi)$. In other words,

$$\begin{aligned}d\Lambda/dt + \eta D^{-1}|(A_1-\Lambda)\psi|^2 &\leqq \eta^{-1}|\varphi\psi|^2|\psi|^{-2} \\ &\leqq C\|\psi\|_1^{n/2}|\psi|^{2-n/2}|\psi|^{-2}\|\varphi\|_1^{n/2}|\varphi|^{2-n/2} \\ &\leqq C\|\varphi\|_1^{n/2}|\varphi|^{2-n/2}\Lambda^{n/4},\end{aligned}$$

and the coefficient on the right-hand side is exponentially small. The existence of the finite limit of $\Lambda(t)$ then follows easily, and the compactness of A_1^{-1} guarantees that the limit of $\Lambda(t)$ is an eigenvalue of A_1. For details of similar calculations refer to [10], [4], and [5].

Remark 6.2. We may also treat in a similar manner the generalized (higher-order) Yukawa interaction model considered in [2] and in [9], i.e., the system (6.10), (6.11) with the coupling terms $-p|\psi|^{2p-2}\psi\varphi$ and $|\psi|^{2p}$, $p \geqq 1$. We do not present here in detail the technically complicated proof of an analogue of Proposition 6.2, which holds under the following restrictions on p: $p < \frac{5}{2}$ for $n=1$, $p < \frac{3}{2}$ for $n=2$, and $p < \frac{7}{6}$ for $n=3$.

7. The linearized flow and differentiability of the nonlinear group. This section deals with some technical tools that will be used to estimate the uniform Lyapunov exponents on the attractor $\mathscr{A}$ constructed in § 4, and hence to prove its finite dimension. We will study the evolution of N-dimensional volumes in $\mathbf{V}$ transported by the differential $DS(t)(\langle\psi_0, \varphi_0, \varphi_1\rangle)$ of the group $S(t)$. First it is necessary to collect some estimates of the system (0.1), (0.2) linearized about the solution $\langle\psi(t), \varphi(t), \varphi_t(t)\rangle$ passing through $\langle\psi_0, \varphi_0, \varphi_1\rangle \in \mathscr{A}$ at $t=0$, i.e., of the equations in variations

$$iz_t - A_1 z + i\varepsilon z + \psi u + \varphi z = 0, \tag{7.1}$$

$$u_{tt} + \delta u_t + A_2 u = \bar{\psi} z + \psi\bar{z}. \tag{7.2}$$

This system is obtained differentiating formally (0.1), (0.2) with respect to $\langle\psi, \varphi\rangle$. It is quite easy to check that this nonautonomous linear system has a unique solution $\langle z, u, u_t\rangle \in C(\mathbb{R}, \mathbf{V})$ if $\langle z_0, u_0, u_1\rangle \in \mathbf{V}$.

We expect that (at least higher modes of) the solutions would be exponentially damped, so we introduce new dependent variables $Z(t) = \exp(\sigma t)z(t)$, $U(t) = \exp(\sigma t)u(t)$ with some small $\sigma > 0$. The system (7.1), (7.2) is rewritten now as

$$iZ_t - A_1 Z + i(\varepsilon - \sigma)Z + \psi U + \varphi Z = 0, \tag{7.3}$$

$$U_{tt} + (\delta - 2\sigma)U_t + (A_2 - \sigma(\delta - \sigma))U = \bar{\psi} Z + \psi\bar{Z}. \tag{7.4}$$

We derive some energy equations and inequalities proceeding as in § 2. The multipliers used to obtain them correspond to "linearized multipliers" in § 2. For (7.3) we have

$$\frac{d|Z|^2}{dt} = -2(\varepsilon-\sigma)|Z|^2 - 2\,\mathrm{Im}\,(\psi U, Z), \tag{7.5}$$

$$\frac{d(A_1Z, Z)}{dt} = -2(\varepsilon-\sigma)(A_1Z, Z) + 2\,\mathrm{Im}\,(A_1Z, \psi U) + 2\,\mathrm{Im}\,(A_1Z, \varphi Z), \tag{7.6}$$

$$\begin{aligned}\frac{d(\mathrm{Re}\,(\psi U, Z))}{dt} &= \mathrm{Re}\,(\psi_t U, Z) + \mathrm{Re}\,(\psi U_t, Z) - (\varepsilon-\sigma)\,\mathrm{Re}\,(\psi U, Z)\\ &\quad - \mathrm{Im}\,(\psi U, A_1Z) + \mathrm{Im}\,(\psi U, \varphi Z),\end{aligned} \tag{7.7}$$

$$\frac{d(\varphi Z, Z)}{dt} = (\varphi_t Z, Z) - 2\,\mathrm{Im}\,(\varphi Z, A_1Z) - 2(\varepsilon-\sigma)(\varphi Z, Z) + 2\,\mathrm{Im}\,(\varphi Z, \psi U). \tag{7.8}$$

For (7.4) we obtain

$$\frac{1}{2}\frac{d(|U_t|^2 + ((A_2 - \sigma(\delta-\sigma))U, U))}{dt} + (\delta - 2\sigma)|U_t|^2 = 2\,\mathrm{Re}\,(\psi U_t, Z), \tag{7.9}$$

$$\frac{d(U_t, U)}{dt} - |U_t|^2 + \left(\frac{1}{2}\delta - \sigma\right)\frac{d|U|^2}{dt} + ((A_2 - \sigma(\delta-\sigma))U, U) = 2\,\mathrm{Re}\,(\psi U, Z). \tag{7.10}$$

A linear combination of (7.5)-(7.10) gives us

$$\begin{aligned}&\frac{dq}{dt} + 2(\varepsilon-\sigma)(A_1Z, Z) + 2\nu(\varepsilon-\sigma)|Z|^2 + (\delta - 2\sigma - \mu)|U_t|^2 + \mu(A_2U, U)\\ &= \mu\sigma(\delta-\sigma)|U|^2 + (\sigma\delta - \sigma^2 + \mu^2 - \mu\sigma + 2\mu\sigma)(U_t, U) - 2\,\mathrm{Re}\,(\psi_t U, Z)\\ &\quad + 2(\varepsilon - \sigma + \mu)\,\mathrm{Re}\,(\psi U, Z) - (\varphi_t Z, Z) + 2(\varepsilon-\sigma)(\varphi Z, Z) - 2\nu\,\mathrm{Im}\,(\psi U, Z),\end{aligned} \tag{7.11}$$

where

$$\begin{aligned}q = q(Z, U, U_t) &= (A_1Z, Z) - 2\,\mathrm{Re}\,(\psi U, Z) - (\varphi Z, Z) + \nu|Z|^2\\ &\quad + \tfrac{1}{2}|U_t + \mu U|^2 + \tfrac{1}{2}(A_2U, U).\end{aligned} \tag{7.12}$$

Taking sufficiently small $\mu = \sigma > 0$ we can choose sufficiently large $\nu > 0$ to finally get

$$\frac{dq}{dt} \leqq C\nu|U||Z|. \tag{7.13}$$

The estimates necessary to do so are collected here:

$$\begin{aligned}&2|(U_t, U)| \leqq |U_t|^2 + |U|^2,\\ &|(\psi_t U, Z)| \leqq |\psi_t||U|_4|Z|_4 \leqq C\|U\|_1^{3/4}|U|^{1/4}\|Z\|_1^{3/4}|Z|^{1/4}\\ &\qquad \leqq \tfrac{1}{4}\mu(A_2U, U) + \tfrac{1}{4}\varepsilon(A_1Z, Z) + C|U||Z|,\\ &|(\psi U, Z)| \leqq |\psi|_\infty|U||Z| \leqq C|U||Z|,\\ &|(\varphi_t Z, Z)| \leqq |\varphi_t|_6|Z|_{12/5}^2 \leqq C\|\varphi_t\|_1\|Z\|_1^{1/2}|Z|^{3/2} \leqq \tfrac{1}{2}\varepsilon(A_1Z, Z) + C|Z|^2,\\ &|(\varphi Z, Z)| \leqq |\varphi|_\infty|Z|^2 \leqq C|Z|^2.\end{aligned}$$

We have used (1.2), (1.3) above and the boundedness of the attractor $\mathscr{A}$ in $\mathbf{W}$.

Remark 7.1. The importance of the quantity q defined in (7.12) follows from the fact that for sufficiently large ν and small μ, $q^{1/2}$ is equivalent to the norm in $\mathbf{V}$. This is a consequence of the Cauchy-Schwarz inequality applied to (7.12) and L^∞ uniform bounds for ψ and φ.

Now we sketch a proof of the uniform differentiability in $\mathbf{V}$ of the mappings $S(t)$ restricted to invariant sets bounded in $\mathbf{W}$, that is, we show the following proposition.

PROPOSITION 7.1. *Let X be a bounded set in $\mathbf{W}$. Then*

$$\lim_{r\to 0} \sup_{\substack{\langle\psi_0^{(k)},\varphi_0^{(k)},\varphi_1^{(k)}\rangle\in X,\ k=1,2,\\ |\langle\Psi_0,\Phi_0,\Phi_1\rangle|_V\leqq r}} |\langle\Psi_0,\Phi_0,\Phi_1\rangle|_V^{-1}|S(t)(\langle\psi_0^{(2)},\varphi_0^{(2)},\varphi_1^{(2)}\rangle),$$

$$-S(t)(\langle\psi_0^{(1)},\varphi_0^{(1)},\varphi_0^{(1)}\rangle)+DS(t)(\langle\psi_0^{(1)},\varphi_0^{(1)},\varphi_1^{(1)}\rangle)(\langle\Psi_0,\Phi_0,\Phi_1\rangle)|_V=0,$$

locally uniformly in t.

The notation used here is that of § 4: $\Psi=\psi^{(1)}-\psi^{(2)}$, $\Phi=\varphi^{(1)}-\varphi^{(2)}$.

This property will be used with $X=\mathscr{A}$. This technical assumption is used in general theorems proved in [7] and [14] on the finite dimension of $S(t)$ invariant sets that are compact in $\mathbf{V}$.

Proof. The functions Ψ, Φ satisfy system (4.2), (4.3). We derive from the energy identity (4.8) with $B_1=A_1$ the following inequality (remember that now $\psi^{(k)}$, $\varphi^{(k)}$, $k=1,2$, are uniformly bounded in H^2):

$$\frac{d(A_1\Psi,\Psi)}{dt}+2\varepsilon(A_1\Psi,\Psi)\leqq C(A_1\Psi,\Psi)^{1/2}((A_1\Psi,\Psi)^{1/2}+(A_2\Phi,\Phi)^{1/2}). \tag{7.14}$$

We used above the elementary inequality

$$\begin{aligned}\|wv\|_1^2=|wv|^2+|\nabla(wv)|^2&\leqq\|w\|_1^2\|v\|_1^2+|w|_\infty^2\|v\|_1^2+|\nabla w|_4^2|v|_4^2\\ &\leqq C\|w\|_2^2\|v\|_1^2\quad\text{valid for all } w\in H^2,\quad v\in H^1.\end{aligned} \tag{7.15}$$

Multiplying (4.3) by Φ_t, we get

$$\begin{aligned}\frac{d(\frac{1}{2}|\Phi_t|^2+\frac{1}{2}(A_2\Phi,\Phi))}{dt}+\delta|\Phi_t|^2&=(\psi^{(1)}\Phi_t,\Psi)+(\Psi\Phi_t,\psi^{(2)})\\ &\leqq C|\Phi_t|(A_1\Psi,\Psi)^{1/2}.\end{aligned} \tag{7.16}$$

As a consequence of the sum of (7.14), (7.16) we obtain

$$|\langle\Psi(t),\Phi(t),\Phi_t(t)\rangle|_V\leqq\exp(Ct)|\langle\Psi(0),\Phi(0),\Phi_t(0)\rangle|_V \tag{7.17}$$

with a constant C. This clearly gives the continuity of $S(t)$ in $\mathbf{V}$ when restricted to bounded subsets of $\mathbf{W}$.

We now consider the solution $\langle z,u\rangle$ of the system of equations in variations (7.1), (7.2) with the initial conditions $z(0)=-\Psi(0)$, $u(0)=-\Phi(0)$, $u_t(0)=-\Phi_t(0)$. Our aim is to estimate the functions $\xi=\Psi+z$ and $\zeta=\Phi+u$ in the H^1 norm. They satisfy the system

$$i\xi_t-A_1\xi+i\varepsilon\xi=-\psi^{(1)}\zeta-\varphi^{(1)}\xi+\Psi\Phi, \tag{7.18}$$

$$\zeta_{tt}+\delta\zeta_t+A_2\zeta=\overline{\psi^{(1)}}\xi+\psi^{(1)}\bar{\xi}-|\Psi|^2, \tag{7.19}$$

following from (4.2), (4.3) and (7.1), (7.2); moreover, $\xi(0)=0$, $\zeta(0)=0$, $\zeta_t(0)=0$.

The associated energy inequalities are

$$\frac{d(A_1\xi,\xi)}{dt}+2\varepsilon(A_1\xi,\xi)=2\,\mathrm{Im}\,(A_1^{1/2}\xi, A_1^{1/2}(\varphi^{(1)}\xi+\psi^{(1)}\zeta-\Psi\Phi)) \tag{7.20}$$

$$\leqq C(A_1\xi,\xi)^{1/2}(\|\psi^{(1)}\xi\|_1+\|\psi^{(2)}\zeta\|_1+\|\Psi\Phi\|_1),$$

$$\frac{1}{2}\frac{d(|\zeta_t|^2+(A_2\zeta,\zeta))}{dt}+\delta|\zeta_t|^2=(\psi^{(1)}\zeta_t,\xi)+(\xi,\psi^{(1)}\zeta_t)-(\zeta_t,|\Psi|^2) \tag{7.21}$$

$$\leqq C(|\zeta_t||\xi|+|\xi||\zeta_t|+|\zeta_t|^2+\|\Psi^2\|_1^2).$$

Combining (7.20) and (7.21) and using (7.15) several times, we arrive at

$$\frac{dJ}{dt}\leqq C(J+\|\Psi\Phi\|_1^2+\|\Psi^2\|_1^2), \tag{7.22}$$

where $J=(A_1\xi,\xi)+(A_2\zeta,\zeta)+|\zeta_t|^2$, $J(0)=0$. If we modify (7.15) slightly to obtain $\|wv\|_1^2\leqq C(\|w\|_1^2+\|w\|_2^{3/2}\|w\|_1^{1/2})\|v\|_1^2$, then (7.22) will imply $dJ/dt\leqq C(J+\|\Psi\|_1^{1/2}(\|\Phi\|_1^2+\|\Psi\|_1^2))$ (remember that the norms $\|\Psi\|_2$, $\|\Phi\|_2$ are uniformly bounded). After the integration this gives $J(t)\leqq C(t)|\langle\Psi(0),\Phi(0),\Phi_t(0)\rangle|_V^{5/2}$, which concludes the proof of Proposition 7.1.

8. Dimension of compact invariant sets. In this section we will prove the main result of this paper, that every $S(t)$ invariant set $X\subset\mathbf{V}$ bounded in $\mathbf{W}$ is finite-dimensional. Here the dimension is understood as the fractal (or entropy) dimension calculated with respect to the norm in $\mathbf{V}$. This result implies, of course, that the Hausdorff dimension of such X is also finite. In particular, X is homeomorphic to a subset of Euclidean space. Since we can take as X the global attractor $\mathscr{A}$ constructed in § 4, this result may be interpreted as the asymptotically finite-dimensional character of dynamics of the system (0.1), (0.2); this dynamics is governed by a finite number of parameters when $t\to+\infty$; the system has a finite number of degrees of freedom when $t\to+\infty$. The proof is based on the use of classical tools from [7], modified in [14], to work with noncompact mappings $S(t)$, and some new technical ingredients from [11] and [12]. These last improvements consist of the use for calculating the dimension of an equivalent norm in $\mathbf{V}$, which is related to intrinsic energies of the system more closely than the original norm. Our § 7 contains the proof of the important part of assumptions of the abstract theorem in [7] ((3.50), (3.51) in [11]) as well as some tools for calculating the evolution of N-dimensional volumes carried by the differential $DS(t)(\cdot)$ of the flow.

We study the volumes of the N-dimensional polyhedra spanned by the vectors $v_0^{(1)},\cdots,v_0^{(N)}\in\mathbf{V}$ and their evolution in time under the action of $DS(t)(\langle\psi_0,\varphi_0,\varphi_1\rangle)$, or more precisely the Gram determinants

$$\det_{1\leqq j,k\leqq N}(v^{(j)}(t),v^{(k)}(t))=|v^{(1)}(t)\wedge\cdots\wedge v^{(N)}(t)|_V^2, \tag{8.1}$$

where $v^{(j)}(t)=DS(t)(\langle\psi_0,\varphi_0,\varphi_1\rangle)(v_0^{(j)})$ with $v_0^{(j)}=\langle z_0^{(j)},u_0^{(j)},u_1^{(j)}\rangle\in\mathbf{V}$. These determinants are the squares of $N!\times$volumes of these polyhedra. We now prove Proposition 8.1.

PROPOSITION 8.1. *For every $S(t)$ invariant set X bounded in $\mathbf{W}$ there exist the constants $\sigma>0$, $\alpha>0$, and C, C_0 such that for every element $\langle\psi_0,\varphi_0,\varphi_1\rangle$ of X, $N\in\mathbb{N}$ and $T\geqq 0$*

$$|v^{(1)}(t)\wedge\cdots\wedge v^{(N)}(t)|\leqq C^N\exp((C_0N^{1-\alpha}-\sigma N)t)|v_0^{(1)}\wedge\cdots\wedge v_0^{(N)}|,$$

for all $v_0^{(j)}\in\mathbf{V}$, $j=1,\cdots,N$, $0\leqq t\leqq T$.

Consequently, for sufficiently large Nt the N-dimensional volumes transported by the tangent flow $DS(t)(\cdot)$ are uniformly contracted.

Proof. As in § 7 it is more convenient to work with new variables $w^{(j)}(t) = \exp(\sigma t)v^{(j)}(t)$ with $\sigma > 0$ determined in § 7. Here v is a short notation for $\langle z, u, u_t\rangle$ and w corresponds to $\langle Z, U, U_t\rangle$ used before in (7.1)-(7.4). Therefore

$$|v^{(1)}(t) \wedge \cdots \wedge v^{(N)}(t)|_V^2 = \exp(-2\sigma N t) G_N(t), \tag{8.2}$$

where $G_N(t) = |w^{(1)}(t) \wedge \cdots \wedge w^{(N)}(t)|_V^2$. As we remarked before, the norm in $\mathbf{V}$ is not particularly well suited for studying the evolution of $G_N(t)$. The expression in (7.12) that we recall here

$$\begin{aligned} q(Z, U, U_t) = (A_1 Z, Z) - 2\,\mathrm{Re}\,(\psi U, Z) - (\varphi Z, Z) \\ + \nu|Z|^2 + \tfrac{1}{2}|U_t + \mu U|^2 + \tfrac{1}{2}(A_2 U, U) \end{aligned} \tag{8.3}$$

is much better in view of (7.13). Furthermore, we consider the $\mathbb{R}$-bilinear form Q on the real copy of $\mathbf{V}\times\mathbf{V}$ that is defined as the polarization of the quadratic form q on $\mathbf{V}$ and

$$H_N(t) = \det_{1 \leqq j,k \leqq N} Q(w^{(j)}(t), w^{(k)}(t)), \tag{8.4}$$

the Gram determinant with respect to the form Q. The fact of crucial importance here is the equivalence of $q^{1/2}$ (with suitably large ν) and the usual norm on $\mathbf{V}$ calculated along the trajectories for each t:

$$C^{-1}|\langle Z, U, U_t\rangle|_V^2 \leqq q(Z, U, U_t) \leqq C|\langle Z, U, U_t\rangle|_V^2 \tag{8.5}$$

for some positive constant C independent of time (see Remark 7.1). Therefore Q is bounded and coercive, and hence defines a scalar product on $\mathbf{V}\times\mathbf{V}$. This implies the equivalence of the Gram determinants $G_N(t)$ and $H_N(t)$ for each t:

$$C^{-N} G_N(t) \leqq H_N(t) \leqq C^N G_N(t) \tag{8.6}$$

(see Lemma 1 of the Appendix in [12]). Inequality (7.13) reads (we use the notation consistent with that of the Appendix in [12])

$$\frac{dq(Z, U, U_t)}{dt} \leqq C|U||Z| \leqq C(K\langle Z, U, U_t\rangle, \langle Z, U, U_t\rangle)_V, \tag{8.7}$$

where $K : \mathbf{V} \to \mathbf{V}$ is defined by

$$K\langle Z, U, U_t\rangle = \langle A_1^{-1} Z, A_2^{-1} U, 0\rangle. \tag{8.8}$$

Obviously, K is a compact operator in $\mathbf{V}$ and its spectrum is contained in the union of spectra of A_1^{-1}, A_2^{-1}, i.e., in the union of two sequences $((\lambda_j^{(k)})^{-1})$ decreasing to zero and (0), $j \in \mathbb{N}$, $k = 1, 2$. The multiplicity of the eigenvalue $(\lambda_j)^{-1}$ of K is the sum of the multiplicities of the eigenvalues of A_1 and A_2 coinciding with λ_j. Theorem A of the Appendix in [12] implies the inequality

$$G_N(t) \leqq C^N \exp\left(\left(C_0 \sum_{j=1}^{N} \lambda_j^{-1}\right) t\right) G_N(0), \tag{8.9}$$

for all $0 \leqq t \leqq T$ and $N \in \mathbb{N}$. The proof of this theorem heavily uses (8.6), (8.7).

Finally, the classical Weyl formula for the asymptotics of the eigenvalues of second-order elliptic operators permits us to estimate the order of the sum $s_N = \sum_{j=1}^{N} \lambda_j^{-1}$ for $n = 1, 2, 3$:

$$s_N \leqq C \quad \text{is less than a constant,} \tag{8.10.1}$$

$$s_N \leqq C \log N \quad \text{as} \sum_{j=1}^{N} j^{-2/n} \simeq \log N, \tag{8.10.2}$$

$$s_N \leqq CN^{1/3} \quad \simeq N^{1/3}. \tag{8.10.3}$$

Inequalities (8.9) and (8.3) conclude the proof of Proposition 8.1. Therefore, taking N such that $CN^{1/3} - \sigma N < 0$ (so $\alpha = \frac{2}{3}$ is good), we can apply the general result from [7]. For such N and for sufficiently large t, the N-dimensional volumes in $\mathbf{V}$ are uniformly contracted by $DS(t)(\langle \psi, \varphi, \varphi_t \rangle)$ for all $\langle \psi, \varphi, \varphi_t \rangle \in X$. Hence the fractal dimension of X is finite. Since the global attractor $\mathscr{A}$ is an $S(t)$-invariant bounded subset of $\mathbf{W}$ we get Theorem 8.1.

THEOREM 8.1. *The global attractor $\mathscr{A}$ defined in* (4.11) *is finite-dimensional and its Hausdorff dimension can be estimated from above by* $N+1$.

A bound for the fractal dimension of $\mathscr{A}$ is more complicated; it contains the values of global Lyapunov exponents intimately related with the contraction rates in Proposition 8.1.

A detailed description of relations between Lyapunov exponents and the fractal dimension in the general setting can be found in [7] and in [11, § 3.3].

9. Remarks on the behavior of the perturbed attractor. In this section we give some remarks on the asymptotic behavior of the family of systems related to (0.1), (0.2):

$$i\psi_t - A_1\psi + i\varepsilon\psi = -\psi\varphi + F, \tag{9.1}$$

$$\beta^2\varphi_{tt} + \beta\varphi_t + A_2\varphi = |\psi|^2 + G, \tag{9.2}$$

with $\varepsilon, \beta > 0$. Of course, for fixed $\varepsilon, \beta > 0$ the results of §§ 2-4, 7, and 8 apply in this situation but we are interested in a singular limit $\beta \to 0$, $\varepsilon \to 0$. From the physical point of view, problem (9.1), (9.2) with $F = G = 0$ and $\beta \to 0$ corresponds to the infinite limit of the velocity of propagation of disturbances in the Klein-Gordon equation, and so to an instantaneous response of the field φ to variations of the field ψ. To see this, let us transform the linear part of (9.2) using the slow time $\tau = \beta t$ into $d^2\Phi/d\tau^2 + d\Phi/d\tau + A_2\Phi$, where $\Phi(x, \tau) = \varphi(x, t)$.

A similar problem has been studied in [24] for the Zakharov system in $\mathbb{R}^n$, $n = 1, 2, 3$:

$$\lambda^{-2} n_{tt} - \Delta(n + |E|^2) = 0, \tag{9.3}$$

$$iE_t + \Delta E - nE = 0, \tag{9.4}$$

when $\lambda \to +\infty$. Formally, taking the limit $\beta \to 0^+$ or $\lambda \to +\infty$ uncouples the equations in both systems. As the result of this formal procedure we obtain a Poisson equation $A_2\varphi = |\psi|^2$ on Ω as the limit in (9.2) or $n + |E|^2 = 0$ in $\mathbb{R}^n$ in (9.3). After substitutions we arrive at the Schrödinger equations

$$i\psi_t - A_1\psi = -\psi A_2^{-1}(|\psi|^2), \tag{9.5}$$

$$iE_t + \Delta E + |E|^2 E = 0, \tag{9.6}$$

respectively. The justification of this formal argument was a very difficult problem for the Zakharov system because the solutions of (9.3), (9.4) may develop singularities in finite time and, in fact, the limiting nonlinear cubic Schrödinger equation (9.6) possesses such blowing-up solutions. The theory of existence and regularity of solutions for this

system is still not complete and may require some new techniques (at least for $n=3$).

We will prove the weak convergence of the solutions of (9.1) to those of (9.5) when $\beta \to 0$, $\varepsilon \to 0$, $\varepsilon \simeq \beta$, and we will establish H^1 bounds for $\psi = \psi(\varepsilon, \beta)$ uniformly with respect to $0 < \beta \ll 1$. This will allow us to give a property of the limit behavior of the global attractors $\mathscr{A}(\varepsilon, \beta)$ when $\varepsilon, \beta \to 0$ in $\mathbf{V}$. The questions that remain unanswered now are the following. What is the behavior of $\langle \psi, \varphi, \beta\varphi_t \rangle$ in $\mathbf{W}$ when $\varepsilon, \beta \to 0$? What are the relations between $\mathscr{A}(\varepsilon, \beta)$ and the asymptotic behavior of the solutions to the conservative equation (9.5)?

Several authors have considered problems related to perturbations of attractors of dissipative equations (see, e.g., [16] in a general setting and [17], [22], [23] for the single hyperbolic equation). Let us observe that the mechanism of dissipation introduced in (9.2) combines the effects of increasing frequency of oscillations (at least for the higher modes of φ with $\lambda_k^{(2)} > \frac{1}{4}$) and increasing damping rate β^{-1}, so it is quite different from that for $0 < \delta \ll 1$ in (0.2). It also differs from the damping effects in the singular perturbation problem considered in [17] and [23] and from large viscous damping in the wave equation studied in [22].

We begin with establishing the uniform bounds with respect to β and ε H^1-bounds for $\langle \psi, \varphi, \beta\varphi_t \rangle$.

PROPOSITION 9.1. *Let $\psi = \psi(\varepsilon, \beta)$, $\varphi = \varphi(\varepsilon, \beta)$ be the solutions of the system* (9.1), (9.2) *with the parameters ε, β converging to zero in such a way that the quotient ε/β belongs to a fixed interval* $[1, M]$. *Moreover, suppose that $|F|_2, |F_t|_2 = \mathcal{O}(\varepsilon)$, $|G|_2 = \mathcal{O}(\beta)$, and the initial conditions $\langle \psi_0, \varphi_0, \varphi_1 \rangle$ stay in a bounded subset of* $\mathbf{V}$. *Then*

$$\sup_{\beta,\varepsilon>0} \sup_{t \geq 0} (\|\psi(t; \varepsilon, \beta)\|_1 + \|\varphi(t; \varepsilon, \beta)\|_1 + \beta|\varphi_t(t; \varepsilon, \beta)|_2)$$

is finite.

Proof. The proof follows from a careful analysis of the modified estimates from § 2. The energy equation for (9.2) analogous to (2.9) is

$$\frac{1}{2}\frac{d(\beta^2|\varphi_t + \mu\varphi|^2 + (A_2\varphi, \varphi))}{dt} + \beta(1-\beta\mu)|\varphi_t + \mu\varphi|^2 + \mu(A_2\varphi, \varphi) = (|\psi|^2 + G + \beta\mu(1-\beta\mu)\varphi, \varphi_t + \mu\varphi), \tag{9.7}$$

where $\mu = \frac{1}{2}\beta$ will be posed. Estimating the right-hand side terms in a manner similar to that of (2.11)–(2.12) and using the assumptions on F, F_t, G we get an analogue of (2.13):

$$\frac{dZ}{dt} + \mu Z \leq C(\varepsilon^{-1}|F|^2 + \varepsilon^{-1}|F_t|^2 + \beta^{-1}|G|^2 + \varepsilon^3\mu^{-2}|\psi|^6), \tag{9.8}$$

with $\mu = \frac{1}{2}\beta \leq \frac{1}{2}\varepsilon$, $Z = (A_1\psi, \psi) - (|\psi|^2, \varphi) + 2\,\mathrm{Re}\,(F, \psi) + \frac{1}{2}\beta^2|\varphi_t + \mu\varphi|^2 + (A_2\varphi, \varphi)$, and C independent of t, β, ε.

Inequality (2.7) reads $|\psi(t)|_2 \leq \exp(-\varepsilon t)|\psi_0|_2 + \varepsilon^{-1}|F|$; the rough estimate $|\psi(t)| \leq C$ does not suffice in this case. Inserting this into (9.8), we get

$$Z(t; \varepsilon, \beta) \leq C \exp(-\mu t) + \mathcal{O}(\varepsilon^2\mu^{-2}) + \mathcal{O}(\beta\mu^{-1}). \tag{9.9}$$

This bound for Z is uniform in β, $0 < \beta \ll 1$. Finally, the estimate of Z from below $4Z \geq 4\beta^2|\varphi_t + \mu\varphi|^2 + (A_1\psi, \psi) + (A_2\varphi, \varphi) - C$ similar to that in the proof of Proposition 2.1 concludes the proof of Proposition 9.1.

As the consequence of Proposition 9.1 we have the existence of a universal absorbing set in $V_1 \times V_2$ for all β, $\varepsilon > 0$, $\varepsilon \simeq \beta$. The attractors $\mathscr{A}(\varepsilon, \beta)$ exist for each β, $\varepsilon > 0$ and they are bounded in $\mathbf{W}$ as in Theorem 4.1. However we are not able to prove

either H^2 estimates (as in § 3) independent of β, ε, or an analogue of Proposition 4.3 independent of β, $\varepsilon > 0$, which would show the uniform boundedness in $W_1 \times W_2$ of the attractors $\mathscr{A}(\varepsilon, \beta)$. An analysis of the preceding proofs gives bounds of order β^{-1} for H^2 norms of φ and ψ (observe that in Proposition 4.3 the norm of $P(t, s)$ depends on $|\varphi_t|$; for the details see [20, Thm. 10.1, Chap. 3]).

Finally, we have the convergence (on each finite time interval) of the corresponding individual solutions of the system (9.1), (9.2) to those of (9.5).

PROPOSITION 9.2. *Under the hypotheses of Proposition* 9.1 *and assuming that* $\|\psi_0(\varepsilon, \beta) - \psi_0\|_1 \to 0$, *the difference* $(\varphi(\varepsilon, \beta) - A_2^{-1}(|\psi(\varepsilon, \beta)|^2))$ *converges to zero weak-star in* $L^\infty(\mathbb{R}^+, V_2)$, $\psi(\varepsilon, \beta)$ *converges to a function* ψ *in* $L^p_{\text{loc}}(\mathbb{R}^+, H^{1-a})$ *for any* $a > 0$ *and* $1 < p < \infty$. *The limit function* ψ *is the unique solution in* V_1 *of* (9.5) *with the initial condition* ψ_0.

Proof. The proof repeats the arguments of the demonstration of Theorem 5 of [24] in a slightly simpler situation. Here we indicate only the main steps.

The family of $\psi_t(\varepsilon, \beta)$ is uniformly bounded in $L^\infty_{\text{loc}}(\mathbb{R}^+, V_1')$ as a consequence of Proposition 9.1 and (9.1); hence $\psi(\varepsilon, \beta)$ stay in a bounded subset of $H^1_{\text{loc}}(\mathbb{R}^+, V_1')$. Using the compactness theorems for such vector functions, we obtain weak-star convergence of a subsequence of $\langle \psi(\varepsilon, \beta), \varphi(\varepsilon, \beta) \rangle$ in $L^\infty_{\text{loc}}(\mathbb{R}^+, V_1 \times V_2)$. In particular, $\psi(\varepsilon, \beta)$ is convergent to some ψ in $L^p_{\text{loc}}(\mathbb{R}^+, H^{1-a})$ for any $a > 0$, $1 < p < \infty$. These facts allow us to pass to the limit in the weak formulation of (9.1):

$$\int_{\mathbb{R}^+} \int_\Omega \Big(-i\chi_t \psi - \sum_{i,j} a_{ij}^{(1)} \frac{\partial \chi}{\partial x_i} \frac{\partial \psi}{\partial x_j} + i\varepsilon \chi \psi + \chi \psi A_2^{-1}(|\psi|^2) + \chi \psi (\varphi - A_2^{-1}(|\psi|^2)) - \chi F \Big)\, dx\, dt = \int_\Omega \chi \psi_0\, dx, \tag{9.10}$$

for every $\chi \in C_0^\infty(\mathbb{R}^+ \times \Omega)$. To see that the weak-star limit in $L^\infty(\mathbb{R}^+, V_2)$ of $\varphi(\varepsilon, \beta) - A_2^{-1}(|\psi(\varepsilon, \beta)|^2)$ is zero we consider after Schochet and Weinstein [24] the function $R = R(x, t; \varepsilon, \beta) = \int_0^t \int_0^s (\varphi - A_2^{-1}(|\psi|^2))\, d\tau\, ds$. R is sufficiently regular in t and satisfies the equation

$$\beta^2 R_{tt} + \beta R_t + A_2 R = \beta \varphi(0)(\beta + t) + \beta^2 t \varphi_t(0) - \beta^2 A_2^{-1}(|\psi|^2) - \beta \int_0^t A_2^{-1}(|\psi|^2) + \int_0^t \int_0^s G, \tag{9.11}$$

with the initial conditions $R(0) = 0$, $R_t(0) = 0$. Taking the scalar product of (9.11) with R_t and using the Cauchy-Schwarz inequality we get

$$\frac{1}{2} \frac{d(\beta^2 |R_t|^2 + (A_2 R, R))}{dt} + \beta |R_t|^2 = \mathcal{O}(\beta) + \beta |R_t|^2 + C\beta^{-1} \left| \int_0^t \int_0^s G \right|^2; \tag{9.12}$$

hence $\|R\|_1^2 = \mathcal{O}(\beta)$. This, together with an easy argument showing the uniqueness of solutions to (9.5) in V_1, concludes the proof of Proposition 9.2, since every subsequence of the family $\psi(\varepsilon, \beta)$ converges to the unique ψ.

It would be interesting to reveal relations between the attractors $\mathscr{A}(\varepsilon, \beta)$ of the slightly damped system and the time asymptotics of the solutions of the limit equation. There is a conjecture that $\mathscr{A}(\varepsilon, \beta)$ approximate invariant measures for the flow associated with (9.5). This seems to be a nontrivial problem and requires further study.

Acknowledgments. I thank Jean-Michel Ghidaglia for interesting and fruitful discussions during my stay at the University of Paris-Sud. I also acknowledge the referee's remarks on the first version of this paper.

REFERENCES

[1] A. V. BABIN AND M. I. VISHIK, *Regular attractors of semigroups and evolution equations*, J. Math. Pures Appl. (9), 62 (1983), pp. 441-491.

[2] A. BACHELOT, *Problème de Cauchy pour des systèmes hyperboliques semi-linéaires*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), pp. 453-478.

[3] P. BILER, *Remark on the decay for damped string and beam equations*, Nonlinear Anal., 10 (1986), pp. 839-842.

[4] ———, *Exponential decay of solutions of damped nonlinear hyperbolic equations*, Nonlinear Anal., 11 (1987), pp. 841-849.

[5] ———, *Regular decay of solutions of strongly damped nonlinear hyperbolic equations*, Appl. Anal., 32 (1989), pp. 277-285.

[6] H. BRÉZIS AND T. GALLOUËT, *Nonlinear Schrödinger evolution equations*, Nonlinear Anal., 4 (1980), pp. 677-681.

[7] P. CONSTANTIN, C. FOIAS, AND R. TEMAM, *Attractors representing turbulent flows*, Memoirs Amer. Math. Soc. 53, American Mathematical Society, Providence, RI, 1985.

[8] I. FUKUDA AND M. TSUTSUMI, *On coupled Klein-Gordon-Schrödinger equations* II, J. Math. Anal. Appl., 66 (1978), pp. 358-378.

[9] ———, *On coupled Klein-Gordon-Schrödinger equations* III, Math. Japon., 24 (1979), pp. 307-321.

[10] J.-M. GHIDAGLIA, *Long time behaviour of solutions of abstract inequalities: applications to thermohydraulic and magnetohydrodynamic equations*, J. Differential Equations, 61 (1986), pp. 268-294.

[11] ———, *Finite dimensional behavior for weakly damped driven Schrödinger equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 5 (1988), pp. 365-405.

[12] ———, *Weakly damped forced Korteweg-de Vries equations behave as a finite dimensional dynamical system in the long time*, J. Differential Equations, 74 (1988), pp. 369-390.

[13] J.-M. GHIDAGLIA AND B. HÉRON, *Dimension of the attractors associated to the Ginzburg-Landau partial differential equation*, Phys. D, 28 (1987), pp. 282-304.

[14] J.-M. GHIDAGLIA AND R. TEMAM, *Attractors for damped nonlinear hyperbolic equations*, J. Math. Pures Appl. (9), 66 (1987), pp. 273-319.

[15] J. K. HALE, *Asymptotic behaviour and dynamics in infinite dimensions*, in Nonlinear Differential Equations, J. K. Hale and P. Martínez-Amores, eds., Res. Notes in Math. 132, Pitman, Boston, 1985, pp. 1-42.

[16] J. K. HALE, X.-B. LIN, AND G. RAUGEL, *Upper semicontinuity of attractors for approximations of semigroups and partial differential equations*, Math. Comp., 50 (1988), pp. 89-123.

[17] J. K. HALE AND G. RAUGEL, *Upper semicontinuity of the attractor for a singularly perturbed hyperbolic equation*, J. Differential Equations, 73 (1988), pp. 197-214.

[18] A. HARAUX, *Two remarks on hyperbolic dissipative problems*, in Collège de France, Seminar P.D.E. 1983/84, J. L. Lions and H. Brézis, eds., Res. Notes in Math. 122, Pitman, Boston, 1985, pp. 161-179.

[19] N. HAYASHI AND W. VON WAHL, *On the global strong solutions of coupled Klein-Gordon-Schrödinger equations*, J. Math. Soc. Japan, 39 (1987), pp. 489-497.

[20] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Vols. 1 and 2, Dunod, Paris, 1968.

[21] X. MORA, *Finite dimensional attracting invariant manifolds for damped semilinear wave equation*, in Contributions to Partial Differential Equations II, J. I. Diaz and P. L. Lions, eds., Research Notes in Mathematics 155, Longmans Press, London, 1987, pp. 172-183.

[22] X. MORA AND J. SOLÀ-MORALES, *Existence and non-existence of finite-dimensional globally attracting invariant manifolds in semilinear damped wave equations*, in Dynamics of Infinite Dimensional Systems, S. N. Chow and J. K. Hale, eds., Springer-Verlag, New York, 1987, pp. 187-210.

[23] ———, *The singular limit dynamics of semilinear damped wave equations*, J. Differential Equations, 78 (1989), pp. 262-307.

[24] S. H. SCHOCHET AND M. I. WEINSTEIN, *The nonlinear Schrödinger limit of the Zakharov equations governing Langmuir turbulence*, Comm. Math. Phys., 106 (1986), pp. 569-580.

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1213-1224, September 1990
© 1990 Society for Industrial and Applied Mathematics
006

ON THE LINEAR HEAT EQUATION WITH FADING MEMORY*

ALESSANDRA LUNARDI†

Abstract. The linear heat equation in materials with memory is studied by reducing it to an abstract Volterra equation. Results of regularity, asymptotic behavior, and positivity are given.

Key words. heat equation, fading memory, abstract Volterra equations, completely monotonic functions

AMS(MOS) subject classifications. 35K05, 45K05, 45N05

0. Introduction. In this paper we consider a model for the heat conduction in materials of fading memory type:

$$\begin{aligned} &b_0 u_t(t,x)+\frac{d}{dt}\left(\int_{-\infty}^{t}\beta(t-s)u(s,x)\,ds\right)\\ &\quad=c_0\Delta u(t,x)-\int_{-\infty}^{t}\gamma(t-s)\Delta u(s,x)\,ds+f(t,x), \qquad t\in\mathbb{R},\quad x\in\bar{\Omega},\\ &u(t,x)=0, \qquad t\in\mathbb{R},\quad x\in\partial\Omega \end{aligned} \tag{0.1}$$

where Ω is a bounded open set in $\mathbb{R}^n$ $(n=1,2,3)$, $u(t,x)$ is the temperature of the point $x\in\bar{\Omega}$ at the time $t\in\mathbb{R}$ (which is assumed to vanish at the boundary of Ω), $f(t,x)$ is the heat supply, b_0, c_0 are positive constants, and β, $\gamma:[0,+\infty[\to\mathbb{R}$ are positive, decreasing L^1 functions.

Equation (0.1) has been introduced in [15], whereas nonlinear versions of (0.1) have been formulated in [3]-[6] and [13], and hyperbolic versions may be found in [14] and [11]. In all these papers the history of the temperature is assumed to be known in $]-\infty,0]$, so that (0.1) reduces to

$$\begin{aligned} &b_0 u_t(t,x)+\frac{d}{dt}\left(\int_{-\infty}^{t}\beta(t-s)u(s,x)\,ds\right)\\ &\quad=c_0\Delta u(t,x)-\int_{-\infty}^{t}\gamma(t-s)\Delta u(s,x)\,ds+h(t,x), \qquad t>0,\quad x\in\bar{\Omega},\\ &u(0,x)=u_0(x), \qquad x\in\bar{\Omega},\\ &u(t,x)=0, \qquad t\geqq 0,\quad x\in\partial\Omega \end{aligned} \tag{0.2}$$

where h is a known function. For physical reasons, we must assume

$$c_0-\int_0^{+\infty}\gamma(s)\,ds>0. \tag{0.3}$$

The relaxation functions β and γ are usually taken as

$$\beta(t)=\sum_{i=1}^{n}\beta_i e^{-b_i t}, \qquad \gamma(t)=\sum_{j=1}^{m}\gamma_j e^{-c_j t} \tag{0.4}$$

with β_i, b_i, γ_j, $c_j>0$. We consider a larger class of kernels, i.e., the completely monotonic ones (for equivalent definitions and properties we refer to [18, Chap. 4]). To study

* Received by the editors May 5, 1987; accepted for publication (in revised form) October 31, 1988.
† Dipartimento di Matematica, Università di Pisa, Via Buonarroti 2, 56100 Pisa, Italy. Present address, Dipartimento di Matematica, Università di Cagliari, Via Ospedale 72, 09124 Cagliari, Italy.

(0.1) and (0.2) we rewrite them as evolution equations in the Banach space X of the continuous functions in $\bar{\Omega}$. We show that they may be reduced, respectively (setting $u(t)=u(t,\cdot)$, $f(t)=f(t,\cdot)$, $h(t)=h(t,\cdot)$), to

$$u'(t)=Au(t)+\int_{-\infty}^{t}K(t-s)u(s)\,ds+f(t),\qquad t\in\mathbb{R}, \tag{0.5}$$

$$\begin{aligned}&u'(t)=Au(t)+\int_{0}^{t}K(t-s)u(s)\,ds+h(t),\qquad t>0,\\ &u(0)=u_0\end{aligned} \tag{0.6}$$

where $A: D(A)\subset X\to X$ generates an analytic semigroup in X, and $K(t)$ is a linear continuous operator from $D(A)$ to X, for every t. Using the abstract theory developed in [9], we prove several existence, regularity, and asymptotic behavior results for the solutions of (0.1) and (0.2). In particular, we show that if the function f is continuous, bounded, and Hölder continuous with respect to time, then (0.1) has a unique bounded solution u; if f is T-periodic with respect to time (respectively, constant) then u is T-periodic with respect to time (respectively, constant); if $f(t,x)$ converges (uniformly in $\bar{\Omega}$) to $f_\infty(x)$ as $t\to+\infty$, then $u(t,x)$ converges (uniformly in $\bar{\Omega}$) to the limiting temperature

$$u_\infty(x)=b_0\left(c_0-\int_0^{+\infty}\gamma(s)\,ds\right)^{-1}g(x),\qquad x\in\bar{\Omega} \tag{0.7}$$

where g is the unique solution of $\Delta g=-f_\infty$, $g_{|\partial\Omega}=0$, and $u_t(t,x)$ converges to zero uniformly in $\bar{\Omega}$. Moreover, we show that if $f(t,x)\geqq 0$ for every t and x, then $u(t,x)$ is nonnegative for every t and x, provided that

$$t\to\beta(t)\left(c_0-\int_0^{t}\gamma(s)\,ds\right)^{-1}\text{ is nonincreasing.} \tag{0.8}$$

We obtain similar results also for problem (0.2). Existence, uniqueness, and regularity properties follow easily from the abstract results of [10], whereas asymptotic behavior and positivity are new and need careful study. In particular, the positivity theorem is proved using the properties of completely monotonic functions and the fact that the solution ϕ of

$$\begin{aligned}&\xi\phi(x)-\Delta\phi(x)=\psi(x),\qquad x\in\Omega,\\ &\xi(x)=0,\qquad x\in\partial\Omega\end{aligned}$$

is nonnegative, provided $\xi\geqq 0$ and $\psi\in C(\bar{\Omega})$, $\psi(x)\geqq 0$ for all x. Completely monotonic functions have been previously used in [8] to study the positivity of the elementary solutions to a certain class of hyperbolic partial differential equations.

Our results concerning problem (0.2) are comparable to those of [4], where a nonlinear version of (0.2) has been studied, as an application of the theory (developed in [1]-[3]) of abstract nonlinear Volterra equations with completely positive kernels. The authors choose $\Omega=]0,1[$ and work in the Hilbert space $X=L^2(0,1)$, so that they find L^2-regularity results for the solution u. They state also a positivity preserving result and show that if $h(t,x)$ converges to $h_\infty(x)$ (in $L^2(0,1)$) as $t\to+\infty$, then $u(t,x)$ converges (in $L^2(0,1)$) to the limiting temperature $u_\infty(x)$ defined in (0.7) (where g is the unique solution of $g''(x)=-h_\infty(x)$, $0<x<1$, $g(0)=g(1)=0$). The kernels β and γ are assumed to be completely positive. Completely monotonic kernels are also completely positive; on the other hand, in [4] it is assumed that $\beta'(t)+(\gamma(0)/c_0)\beta(t)\leqq 0$

almost everywhere for $t>0$ (which is stronger than (0.8)). Therefore the class of kernels considered here is not contained in, nor does it contain, that of [4].

The paper is organized as follows: In § 1 we state existence, regularity, and asymptotic behavior results for the abstract problems (0.5) and (0.6), whereas in § 2 we consider the cone-preserving property. In § 3 we transform problems (0.1) and (0.2) to the forms (0.5) and (0.6), respectively, and we apply the results found in the previous sections.

1. Abstract Volterra equations: Existence and asymptotic behavior results. In this section we recall and develop some results (contained in [9], [12], and [10]) concerning a class of integrodifferential equations in general Banach space X:

$$v'(t)=Av(t)+\int_0^t K(t-s)v(s)\,ds+h(t), \qquad t>0, \tag{1.1}$$

$$u'(t)=Au(t)+\int_{-\infty}^t K(t-s)u(s)\,ds+f(t), \qquad t\in\mathbb{R} \tag{1.2}$$

where $A: D(A)\subset X\to X$ is a linear operator such that

$$\begin{aligned}&\exists M>0,\ \omega\in\mathbb{R},\ \theta\in\,]\pi/2,\pi]\text{ such that the resolvent set}\\ &\rho(A)\text{ of }A\text{ contains the sector }S=\{\lambda\in\mathbb{C};\ \lambda\neq\omega,\ |\arg(\lambda-\omega)|<\theta\}\text{ and}\\ &\|\lambda(\lambda-A)^{-1}\|_{L(X)}\leqq M\text{ for }\lambda\in S.\end{aligned} \tag{1.3}$$

Then A generates an analytic semigroup e^{tA} in X (for analytic semigroups in the case of nondense domain, see [16]).

The operator kernel $K(\cdot)$ belongs to $L^1(0,+\infty;\ L(D(A),X))$ ($D(A)$ is endowed with the graph norm); moreover, we assume that the Laplace transform of $K(\cdot)x$ is analytically extendible to S for every $x\in D(A)$, and there are $N,\ \alpha>0$ such that the extension (denoted by $\hat{K}(\lambda)x$) satisfies

$$|\lambda|^{\alpha}\|\hat{K}(\lambda)x\|\leqq N\|x\|_{D(A)}, \qquad \lambda\in S,\quad x\in D(A) \tag{1.4}$$

where $\|\cdot\|$ is the norm in X.

Then there exists a resolvent operator $R(t)$ ($t\geqq 0$) such that $t\to R(t)$ is analytic in $]0,+\infty[$ with values in $L(X,D(A))$, and

$$\begin{aligned}&R'(t)=AR(t)+\int_0^t K(t-s)R(s)\,ds, \qquad t>0,\\ &\lim_{t\to 0^+}R(t)x=x \qquad \forall x\in\overline{D(A)}\end{aligned} \tag{1.5}$$

(by $\overline{D(A)}$ we denote the closure of $D(A)$ in X). $R(t)$ may be represented by the Dunford integral:

$$\begin{aligned}&R(t)\doteq\frac{1}{2\pi i}\int_\gamma e^{\lambda t}(\lambda-A-\hat{K}(\lambda))^{-1}\,d\lambda, \qquad t>0,\\ &R(0)\doteq 1\end{aligned} \tag{1.6}$$

where γ is the curve $\{\lambda\in\mathbb{C};\ |\arg(\lambda-r)|=\theta\}$ (oriented counterclockwise) with r so large that $(\lambda-A-\hat{K}(\lambda))^{-1}$ exists for $\lambda\in\gamma$.

To state precise estimates for $R(t)$ we need some other notation. We fix a maximal domain Ω of analyticity of $\hat{K}(\cdot)$ (as a function with values in $L(D(A),X)$), and we set

$$\rho_0(A,K)\doteq\{\lambda\in\mathbb{C};\ (\lambda-A-\hat{K}(\lambda))^{-1}\text{ exists in }L(X)\}. \tag{1.7}$$

Since $\rho_0(A, K)$ is open, it is not difficult to see (using the Cauchy formula) that for every $\lambda \in \rho_0(A, K)$, $(\lambda - A - \hat{K}(\lambda))^{-1}$ belongs to $L(X, D(A))$, and $\lambda \to (\lambda - A - \hat{K}(\lambda))^{-1}$ is holomorphic in $\rho_0(A, K)$ with values in $L(X, D(A))$. We define an analytic extension of $(\lambda - A - \hat{K}(\lambda))^{-1}$ in the set

$$\rho(A, K) \doteq \rho_0(A, K) \cup \{\text{removable singularities of } \lambda \to (\lambda - A - \hat{K}(\lambda))^{-1}\} \tag{1.8}$$

setting

$$\begin{aligned} &F(\lambda) \doteq (\lambda - A - \hat{K}(\lambda))^{-1} \quad \text{if } \lambda \in \rho_0(A, K), \\ &F(\lambda_0) \doteq \lim_{\lambda \to \lambda_0} (\lambda - A - \hat{K}(\lambda))^{-1} \quad \text{if } \lambda \in \rho(A, K) \backslash \rho_0(A, K) \end{aligned} \tag{1.9}$$

where the limit is in the $L(X, D(A))$ topology. In contrast to the nonintegral case $K \equiv 0$, it can be shown that if λ_0 belongs to $\rho(A, K) \backslash \rho_0(A, K)$, then $F(\lambda_0)$ is not invertible. We denote by $\sigma(A, K)$ the complementary set $\mathbb{C} \backslash \rho(A, K)$, and finally we set

$$\omega(A, K) \doteq \sup \{\operatorname{Re} \lambda;\ \lambda \in \sigma(A, K)\}. \tag{1.10}$$

Then for every $\varepsilon > 0$ there is $M(\varepsilon) > 0$ such that

$$\|R(t)\|_{L(X)} + \|tR'(t)\|_{L(X)} + \|tAR(t)\|_{L(X)} \leqq M(\varepsilon)\, e^{(\omega(A,K)+\varepsilon)t} \quad \forall t > 0. \tag{1.11}$$

From these estimates and (1.5), existence, uniqueness, and several regularity results for the initial value problem for (1.1) have been proved in [9] and [12]. We mention one of these results below.

PROPOSITION 1.1. *Let* (1.3), (1.4) *hold, let x belong to the closure $\overline{D(A)}$ of the domain of A, and let $h : [0, +\infty[\to X$ be locally α-Hölder continuous. Then the function*

$$v(t) \doteq R(t)x + \int_0^t R(t-s)h(s)\, ds, \qquad t \geqq 0 \tag{1.12}$$

belongs to $C([0, +\infty[; X) \cap C^{1,\alpha}_{\mathrm{loc}}(]0, +\infty[; X) \cap C^{\alpha}_{\mathrm{loc}}(]0, +\infty[; D(A))$ and it is the unique solution of (1.1) *such that $\lim_{t \to 0} v(t) = x$.*

Since, in the application to the heat equation in materials with memory, $\sigma(A, K)$ is contained in $\{\lambda \in \mathbb{C};\ \operatorname{Re} \lambda < 0\}$, to simplify notations and statements we assume from now on that

$$\omega(A, K) < 0. \tag{1.13}$$

To give an asymptotic behavior result, we introduce the class of α-Hölder continuous functions in $[0, +\infty[$:

$$\begin{aligned} C^{\alpha}([0, +\infty[; X) \doteq \Big\{ &h : [0, +\infty[\to X;\ \|h\|_{C^{\alpha}([0, +\infty[; X)} \\ &= \sup_{t \geqq 0} \|h(t)\| + \sup_{0 \leqq r < s} \|h(s) - h(r)\|(s-r)^{-\alpha} < +\infty \Big\}. \end{aligned} \tag{1.14}$$

PROPOSITION 1.2. *Let* (1.3), (1.4), *and* (1.13) *hold, let x belong to $\overline{D(A)}$, and let h belong to $C^{\alpha}([0, +\infty[; X)$ $(0 < \alpha < 1)$ be such that*

$$\lim_{t \to +\infty} h(t) = h_{\infty}.$$

Then

(i) $\lim_{t \to +\infty} v(t) = F(0)h_{\infty}$,

(ii) $\lim_{t \to +\infty} Av(t) = AF(0)h_{\infty}$,

(iii) $\lim_{t \to +\infty} v'(t) = 0$.

Proof. Using the representation formula (1.12) and estimates (1.11) (with ε so small that $\omega(A, K)+\varepsilon<0$), we can easily see that $v(t)$ converges to $\int_0^{+\infty} R(s)\,ds\,h_\infty$ as t goes to $+\infty$. Since $\omega(A, K)<0$, the curve γ in (1.6) may be replaced by any curve $\tilde{\gamma}$ contained in $\{\lambda\in\mathbb{C};\ \mathrm{Re}\,\lambda<0\}$ and joining $\infty e^{-i\theta}$ to $\infty e^{i\theta}$. We get

$$\int_0^{+\infty} R(s)\,ds=\frac{1}{2\pi i}\int_{\tilde{\gamma}}-\frac{F(\lambda)}{\lambda}\,d\lambda=F(0) \tag{1.15}$$

and (i) is proved. For every $t>1$ we have

$$Av(t)=AR(t)x+A\int_1^t R(s)(h(t-s)-h(t))\,ds+A\int_0^1 R(s)(h(t-s)-h(t))\,ds$$
$$+A\int_0^t R(s)h(t)\,ds=I_1(t)+I_2(t)+I_3(t)+I_4(t).$$

Due to (1.11), $\lim_{t\to+\infty} I_1(t)=\lim_{t\to+\infty} I_2(t)=0$. Let us show that $I_3(t)$ also goes to zero as t goes to $+\infty$: If we fix $\varepsilon>0$, there is $M\geqq 1$ such that $\|h(t-s)-h(t)\|\leqq\varepsilon$ for every $t\geqq M$ and $s\in[0,1]$; on the other hand, we have $\|h(t-s)-h(t)\|\leqq[h]_{C^\alpha([0,+\infty[;X)}s^\alpha$, so that

$$\|h(t-s)-h(t)\|\leqq\varepsilon^{1/2}s^{\alpha/2}([h])_{C^\alpha([0,+\infty[;X)})^{1/2},\qquad t\geqq M,\quad 0\leqq s\leqq 1,$$

which (together with (1.11)) implies $\|I_3(t)\|\leqq\mathrm{const.}\varepsilon^{1/2}$ for $t\geqq M$. Finally, since $\lim_{t\to+\infty} A\int_0^t R(s)\,ds=A\int_0^{+\infty} R(s)\,ds$ in $L(X)$ (see [10]), we get $\lim_{t\to+\infty} I_4(t)=\lim_{t\to+\infty} Av(t)=A\int_0^{+\infty} R(s)\,ds\,h_\infty$. Statement (ii) now follows from (1.15). Finally, for every $x\in X$ we have

$$(A+\hat{K}(0))F(0)x=\lim_{\substack{\lambda\to 0\\ \mathrm{Re}\,\lambda>0}}(A+\hat{K}(\lambda)-\lambda)F(\lambda)x=-x$$

so that, letting $t\to+\infty$ in (1.1), we get $\lim_{t\to+\infty} v'(t)=0$, and (iii) is also proved. □

Equation (1.2) has been treated in [10], assuming that f belongs to any of the following spaces ($0<\alpha<1$, $\omega>0$):

$$C^\alpha(\mathbb{R};X)\doteq\Big\{f:\mathbb{R}\to X;\ \|f\|_{C^\alpha(\mathbb{R};X)}=\sup_{t\in\mathbb{R}}\|f(t)\|$$
$$+\sup_{r<s}\|f(s)-f(r)\|(s-r)^{-\alpha}<+\infty\Big\}, \tag{1.16}$$

$$C^\alpha_\omega(\mathbb{R};X)\doteq\{f:\mathbb{R}\to X;\ t\to e^{-\omega t}f(t)\in C^\alpha(\mathbb{R};X)\}. \tag{1.17}$$

The latter space consists of exponentially decaying functions as $t\to-\infty$, and it is endowed with the norm $\|f\|_{C^\alpha_\omega(\mathbb{R};X)}=\|e^{-\omega\cdot}f(\cdot)\|_{C^\alpha(\mathbb{R};X)}$.

The following proposition holds.

PROPOSITION 1.3. *Let* (1.3), (1.4), *and* (1.13) *hold. Then for every f belonging to $C^\alpha(\mathbb{R};X)$ $(0<\alpha<1)$, problem* (1.2) *has a unique bounded solution u, given by*

$$u(t)=\int_{-\infty}^t R(t-s)f(s)\,ds,\qquad t\in\mathbb{R}. \tag{1.18}$$

Moreover, u, u', and $Au(\cdot)$ belong to $C^\alpha(\mathbb{R};X)$. For every f belonging to $C^\alpha_\omega(\mathbb{R};X)$ $(0<\alpha<1,\ \omega>0)$ the function u defined in (1.18) *is the unique solution of* (1.2) *such that $t\to e^{-\omega t}u(t)$ is bounded. Moreover, u, u', and $Au(\cdot)$ belong to $C^\alpha(\mathbb{R};X)$.*

Concerning the limiting behavior of $u(t)$ as $t\to\pm\infty$, we can prove the next proposition.

PROPOSITION 1.4. *Let* (1.3), (1.4), *and* (1.13) *hold, and let* $f \in C^\alpha(\mathbb{R}; X)$ $(0<\alpha<1)$ *be such that there exists* $\lim_{t\to+\infty} f(t) = f_\infty$ (*respectively,* $\lim_{t\to-\infty} f(t) = f_{-\infty}$). *Then*

(i) $\lim_{t\to+\infty} u(t) = F(0)f_\infty$ (*respectively,* $\lim_{t\to-\infty} u(t) = F(0)f_{-\infty}$),

(ii) $\lim_{t\to+\infty} Au(t) = F(0)f_\infty$ (*respectively,* $\lim_{t\to-\infty} Au(t) = Af(0)f_{-\infty}$),

(iii) $\lim_{t\to+\infty} u'(t) = 0$ (*respectively,* $\lim_{t\to-\infty} u'(t) = 0$).

The proof is the same as in Proposition 1.2, with obvious modifications, and we omit it.

2. The cone-preserving property. Let C be a closed convex cone in X. A necessary and sufficient condition for $R(t)(C) \subset C$ for every positive t is given by the following proposition (see [10] for a proof).

PROPOSITION 2.1. *Let* (1.3) *and* (1.4) *hold, and let* $R(t)$ $(t \geqq 0)$ *be the resolvent operator defined in* (1.6). *Then* $R(t)(C) \subset C$ *for every* $t \geqq 0$ *if and only if there is* $\lambda_0 \geqq \omega(A, K)$ *such that*

$$(-1)^k F^{(k)}(\lambda)(C) \subset C \quad \forall \lambda > \lambda_0, \quad \forall k \in \mathbb{N}. \tag{2.1}$$

Generally, in the applications C is the cone of the nonnegative functions in some functional space X. Then, recalling Bernstein's theorem on completely monotonic (real) functions, the result of Proposition 2.1 is not surprising, since $\lambda \to F(\lambda)$ is the Laplace transform of $t \to R(t)$, as is easy to check. The infinitely many conditions in (2.1) are verified in some applications, including the heat equation in materials with memory (see § 3 below).

Using Propositions 1.1 and 1.2, we easily get the following corollary.

COROLLARY 2.2. *Let* (1.3), (1.4), *and* (2.1) *hold. Then*

(a) *If* x *belongs to* C, *and* $h:[0, +\infty[\to C$ *is locally Hölder continuous, then the solution* v *of* (1.1) *such that* $v(0) = x$ *has values in* C.

(b) *If* $\omega(A, K) < 0$ *and* $f \in C^\alpha(\mathbb{R}; X)$ *is such that* $f(t) \in C$ *for every* $t \in \mathbb{R}$, *then the unique bounded solution of* (1.2) *has values in* C.

Remark 2.3. Existence and uniqueness of a bounded solution of (1.2) may also be treated in the case where $\omega(A, K) > 0$ (see [10]). But, even if $R(t)(C)$ is contained in C, we do not expect that the unique bounded solution of (1.2) has values in C if f does. Consider, for instance, $X = \mathbb{R}$, $C = [0, +\infty[$: Easy computations show that the resolvent operator for the equation

$$v'(t) = v(t) + 5\int_0^t e^{-2(t-s)} v(s)\, ds, \qquad t \in \mathbb{R}$$

is given by

$$R(t)x = (\tfrac{1}{6}e^{-4t} + \tfrac{5}{6}e^{2t})x, \qquad t \in \mathbb{R}, \quad x \in \mathbb{R}$$

so that $R(t)$ maps C into itself, but the unique bounded solution of

$$u'(t) = u(t) + 5\int_{-\infty}^t e^{-2(t-s)} u(s)\, ds + f(t), \qquad t \in \mathbb{R}$$

is

$$u(t) = \frac{1}{6}\int_{-\infty}^t e^{-4(t-s)} f(s)\, ds - \frac{5}{6}\int_t^{+\infty} e^{2(t-s)} f(s)\, ds, \qquad t \in \mathbb{R},$$

which is not necessarily nonnegative for every nonnegative f.

3. The heat equation in materials with memory. We study here in detail a particular class of equations of the type (1.1), (1.2), arising in the study of heat flow in materials of the so-called fading memory type. The model is discussed in [15], and here we recall briefly its derivation.

Let $\bar{\Omega}$ be a bounded body in $\mathbb{R}^n$ ($n = 1, 2, 3$). Denote by $\varepsilon(t, x)$ ($x \in \bar{\Omega}$) the internal energy at the time t, by $f(t, x)$ the heat supply and by $\vec{q}(t, x)$ the heat flux. The energy balance law yields

$$\varepsilon_t(t, x) = -\operatorname{div}' \vec{q}(t, x) + f(t), \qquad t \in \mathbb{R}, \quad x \in \bar{\Omega}. \tag{3.1}$$

Generally, we assume that the energy and the heat flux depend linearly on the temperature $u(t, x)$ and on its gradient, respectively. This assumption leads to the classical heat equation, which describes sufficiently well the evolution of temperature in many kinds of materials. This is not the case of materials of fading memory type, where ε and $\vec{q}$ are assumed to be given by

$$\varepsilon(t, x) = b_0 u(t, x) + \int_{-\infty}^{t} \beta(t-s) u(s, x)\, ds, \qquad t \in \mathbb{R}, \quad x \in \bar{\Omega}, \tag{3.2}$$

$$\vec{q}(t, x) = -c_0 \nabla u(t, x) + \int_{-\infty}^{t} \gamma(t-s) \nabla u(s, x)\, ds, \qquad t \in \mathbb{R}, \quad x \in \bar{\Omega} \tag{3.3}$$

(there are also nonlinear models, where ∇u is replaced by a nonlinear function of ∇u in (3.3)). Here b_0 and c_0 (called, respectively, heat capacity and thermal conductivity constants) are positive, and the functions β and γ are generally chosen as in (0.3). Moreover, it is assumed (see [15] for a physical motivation) that

$$c_0 - \int_0^{+\infty} \gamma(s)\, ds > 0. \tag{3.4}$$

The mappings

$$t \to \alpha(t) \doteq b_0 + \int_0^t \beta(s)\, ds, \qquad t \to c(t) \doteq c_0 - \int_0^t \gamma(s)\, ds \tag{3.5}$$

are called internal energy and heat flux relaxation functions, respectively.

Replacing (3.2) and (3.3) in (3.1), and assuming that the temperature is zero at the boundary of Ω, we get the linear heat equation

$$\begin{aligned} &b_0 u_t(t, x) + \frac{d}{dt} \int_{-\infty}^{t} \beta(t-s) u(s, x)\, ds \\ &= c_0 \Delta u(t, x) - \int_{-\infty}^{t} \gamma(t-s) \Delta u(s, x)\, ds + f(t, x), \qquad t \in \mathbb{R}, \quad x \in \bar{\Omega}, \\ &u(t, x) = 0, \qquad t \in \mathbb{R}, \quad x \in \partial\Omega. \end{aligned} \tag{3.6}$$

In many papers, the history of u is assumed to be known in $]-\infty, 0]$, so that (3.6) is

replaced by

$$b_0 u_t(t,x)+\frac{d}{dt}\int_0^t \beta(t-s)u(s,x)\,ds$$

(3.7)

$$=c_0\Delta u(t,x)-\int_0^t \gamma(t-s)\Delta u(s,x)\,ds+h(t,x), \qquad t>0, \quad x\in\bar\Omega,$$

$$u(0,x)=u_0(x), \qquad x\in\bar\Omega,$$

$$u(t,x)=0, \qquad t\geqq 0, \quad x\in\partial\Omega$$

for suitable $h:[0,+\infty[\to\mathbb{R}$.

We will study (3.6) and (3.7) for a larger class of kernels than the ones given in (0.3): we will consider completely monotonic kernels, i.e., C^∞ functions $\beta, \gamma:]0,+\infty[\to[0,+\infty[$ such that

$$(-1)^k\beta^{(k)}(t)\geqq 0, \qquad (-1)^k\gamma^{(k)}(t)\geqq 0 \quad \forall t>0, \quad \forall k\in\mathbb{N}. \tag{3.8}$$

For (3.6) to make sense, we assume also that

$$\beta, \beta', \gamma\in L^1(0,+\infty). \tag{3.9}$$

Due to Bernstein's equivalence theorem (see [18, Chap. IV]), β and γ may be represented as

$$\beta(t)=\int_0^{+\infty} e^{-t\omega}\,d\nu(\omega), \quad \gamma(t)=\int_0^{+\infty} e^{-t\omega}\,d\mu(\omega), \quad t>0 \tag{3.10}$$

where $\nu, \mu:]0,+\infty[\to\mathbb{R}$ are suitable positive nondecreasing functions. Then (3.9) is equivalent to

$$\int_0^{+\infty}\frac{d\nu(\omega)}{\omega}<+\infty, \quad \int_0^{+\infty} d\nu(\omega)<+\infty, \quad \int_0^{+\infty}\frac{d\mu(\omega)}{\omega}<+\infty. \tag{3.11}$$

To write problems (3.6) and (3.7) in the abstract form (1.1), (1.2), we set

$$\begin{aligned} &X=C(\bar\Omega) \quad \text{(endowed with the sup norm)},\\ &D(A)=\{\phi\in C(\bar\Omega);\ \Delta\phi\in C(\bar\Omega),\ \phi_{|\partial\Omega}=0\},\\ &(A\phi)(x)=\frac{c_0}{b_0}\Delta\phi(x)-\frac{\beta(0)}{b_0}\phi(x). \end{aligned} \tag{3.12}$$

The Laplace operator Δ is in the sense of distributions, and $D(A)$ is endowed with the graph norm of A. If the boundary of Ω is sufficiently smooth, then $A:D(A)\to X$ satisfies (1.3) (see [17]).

We also set

$$(K(t)\phi)(x)=-\frac{\beta'(t)}{b_0}\phi(x)-\frac{\gamma(t)}{b_0}\Delta\phi(x), \qquad t>0, \quad \phi\in D(A), \quad x\in\bar\Omega \tag{3.13}$$

so that (3.6) reduces to (1.2), with $f(t)$ replaced by $f(t,\cdot)/b_0$, and (3.7) reduces to the initial value problem for (1.1), with $h(t)$ replaced by $h(t,\cdot)/b_0$. Due to (3.9), $t\to K(t)$ belongs to $L^1(0,+\infty;L(D(A),X))$. Since the Laplace transform of K is given (for $\lambda>0$) by

$$\hat K(\lambda)\phi=-\frac{1}{b_0}\left(\int_0^{+\infty}\frac{\omega\,d\nu\,(\omega)}{\lambda+\omega}\phi+\int_0^{+\infty}\frac{d\mu\,(\omega)}{\lambda+\omega}\Delta\phi\right), \tag{3.14}$$

then (due to (3.11)) $\lambda \to \hat{K}(\lambda)$ is analytically extendible in $\mathbb{C}\setminus]-\infty, 0]$, and the extension is given again by (3.14). In order for (1.4) to be satisfied, we assume there is $\alpha \in [0, 1[$ such that the extension of $\hat{\gamma}(\lambda)$ to $\mathbb{C}\setminus]-\infty, 0]$ is bounded by const. $|\lambda|^{-\alpha}$, i.e.,

$$\int_0^{+\infty} \frac{d\mu(\omega)}{\omega^{1-\alpha}} < +\infty. \tag{3.15}$$

Then the assumptions of § 1 are satisfied. To apply the results of §§ 1 and 2, we must study the set $\rho(A, K)$ and the function $F(\lambda)$.

PROPOSITION 3.1. *Let* (3.8), (3.9), (3.15), *and* (3.4) *hold. Let A and K be defined by* (3.12) *and* (3.13), *respectively. Then*

$$\rho_0(A, K) \supset \{z \in \mathbb{C};\ \operatorname{Re} z \geqq 0\}, \tag{3.16}$$

$$F(\lambda) = \frac{b_0}{c_0 - \hat{\gamma}(\lambda)} R\left(\frac{\lambda(b_0 + \hat{\beta}(\lambda))}{c_0 - \hat{\gamma}(\lambda)}, \Delta\right) \quad \forall \lambda \geqq 0 \tag{3.17}$$

where, for $\xi \geqq 0$, $R(\xi, \Delta) = (\Delta - \xi)^{-1}$ *is the resolvent of the Laplace operator with Dirichlet boundary condition.*

Proof. Let $\operatorname{Re} \lambda \geqq 0$ and $\phi \in D(A)$: then

$$(\lambda - A - \hat{K}(\lambda))\phi = \left(\lambda + \frac{\lambda\hat{\beta}(\lambda)}{b_0}\right)\phi + \left(\frac{\hat{\gamma}(\lambda)}{b_0} - \frac{c_0}{b_0}\right)\Delta\phi$$

so that (since $\hat{\gamma}(\lambda) - c_0 \neq 0$, due to (3.4)), $\lambda - A - \hat{K}(\lambda)$ is invertible if and only if

$$\frac{\lambda(b_0 + \hat{\beta}(\lambda))}{c_0 - \hat{\gamma}(\lambda)} \neq -\lambda_n \quad \forall n \in \mathbb{N} \tag{3.18}$$

where $\{-\lambda_n;\ n \in \mathbb{N}\} \subset]-\infty, 0[$ is the sequence of the eigenvalues of the Laplace operator with zero boundary condition. Using the representation formulas (3.10), we easily get

$$\operatorname{Re} \lambda(b_0 + \hat{\beta}(\lambda)) \geqq 0.$$

We also get

$$\operatorname{Re}(c_0 - \hat{\gamma}(\lambda)) \geqq c_0 - \int_0^{+\infty} \gamma(t)\, dt > 0.$$

Therefore, since $-\lambda_n < 0$ for every n, the real part of $-\lambda_n(c_0 - \hat{\gamma}(\lambda))$ is negative, and (3.18) holds. Formula (3.17) follows now easily. □

Since $\rho(A, K)$ contains $\rho_0(A, K)$ and it is an open set, then (3.16) implies $\omega(A, K) < 0$. Formula (3.18) yields a cone-preserving condition, as the following proposition shows.

PROPOSITION 3.2. *Let* (3.8), (3.9), (3.15), *and* (3.4) *hold. Let A, K, c be defined by* (3.12), (3.13), *and* (3.5), *respectively, and let C be the cone of nonnegative functions in* X: $C \doteq \{\phi \in C(\bar{\Omega});\ \phi(x) \geqq 0\ \forall x \in \bar{\Omega}\}$. *If*

$$t \to \beta(t)/c(t) \text{ is nonincreasing in } [0, +\infty[\tag{3.19}$$

then

$$R(t)(C) \subset C \quad \forall t \geqq 0.$$

Proof. Due to Proposition 2.1, it is sufficient to show that for every $\lambda > 0$ we have

$$(-1)^k F^{(k)}(\lambda)(C) \subset C, \qquad k \in \mathbb{N}. \tag{3.20}$$

The proof is in two steps: we show first that $\lambda \to (b_0/c_0 - \hat{\gamma}(\lambda))$ is completely monotonic on $]0, +\infty[$, and then that

$$(-1)^k \frac{d^k}{d\lambda^k}\left(R\left(\frac{\lambda(b_0+\hat{\beta}(\lambda))}{c_0-\hat{\gamma}(\lambda)}, \Delta\right)\right)(C) \subset C, \qquad k \in \mathbb{N}. \tag{3.21}$$

Recalling Leibnitz's rule and (3.17), we will prove (3.20).

Due to (3.4) we have $\hat{\gamma}(\lambda)/c_0 \in]0, 1[$ for $\lambda > 0$, so that

$$\frac{b_0}{c_0-\hat{\gamma}(\lambda)} = \frac{b_0}{c_0} \sum_{n=0}^{\infty} \left(\frac{\hat{\gamma}(\lambda)}{c_0}\right)^n.$$

Since $\gamma(t)$ is nonnegative for every t, then $\lambda \to \hat{\gamma}(\lambda)$ is completely monotonic on $]0, +\infty[$, as is $\lambda \to (\hat{\gamma}(\lambda))^n$ for every n (the product of two completely monotonic functions is obviously completely monotonic). Therefore $\lambda \to b_0/(c_0 - \hat{\gamma}(\lambda))$ is completely monotonic.

Let us show (3.21): since

$$g(\lambda) = \frac{\lambda(b_0+\hat{\beta}(\lambda))}{c_0-\hat{\gamma}(\lambda)}, \qquad \lambda > 0$$

is positive, and $R(\xi, \Delta)$ maps C into itself for positive ξ, then (3.21) holds for $n = 0$. Let us show that g' is completely monotonic in $[0, +\infty[$. To this aim it is sufficient to prove that for every $\varepsilon > 0$, the function $\lambda \to f_\varepsilon(\lambda) = (1/g(\lambda))/(\varepsilon + 1/g(\lambda))$ is completely monotonic in $[0, +\infty[$, and to remark that $-(d/d\lambda)f_\varepsilon(\lambda)/\varepsilon$ converges uniformly in a neighborhood of the real positive semi-axis to g' as $\varepsilon \to 0$, so that g' is also completely monotonic.

To show that $f_\varepsilon(\lambda) = \hat{c}(\lambda)/(\varepsilon(b_0+\hat{\beta}(\lambda)) + \hat{c}(\lambda))$ is completely monotonic, consider the equation

$$b_0 h(t) + \int_0^t [\varepsilon\beta(t-s) + c(t-s)]h(s)\, ds = c(t), \qquad t \geqq 0. \tag{3.22}$$

It is easy to see that (3.22) has a unique solution h, which is of class C^∞ and Laplace transformable, with $\hat{h}(\lambda) = f_\varepsilon(\lambda)$ for $\lambda \geqq 0$. Moreover, writing (3.22) in the form

$$\frac{b_0 h(t)}{c(t)} + \int_0^t \frac{\varepsilon\beta(t-s) + c(t-s)}{c(t)} h(s)\, ds = 1, \qquad t \geqq 0 \tag{3.23}$$

and recalling that

$$\frac{d}{dt}\frac{\varepsilon\beta(t-s)+c(t-s)}{c(t)} \leqq 0 \quad \text{for } 0 \leqq s \leqq t$$

due to (3.19) and to the complete monotonicity of $t \to c(t)$ (which implies that c'/c is not decreasing), it is not difficult to see that $h(t) \geqq 0$ for every $t \geqq 0$. Therefore f_ε is the Laplace transform of a positive function, so that it is completely monotonic. Now we have

$$\frac{d}{d\lambda}(R(g(\lambda), \Delta)) = -g'(\lambda)R^2(g(\lambda), \Delta), \qquad \lambda > 0$$

so that $-(d/d\lambda)(R(g(\lambda), \Delta))$ maps C into itself. Now let $n \geqq 1$ and assume by induction that (3.21) holds for every $k \leqq n$. We have

$$\begin{aligned} \frac{d^{n+1}}{d\lambda^{n+1}} R(g(\lambda), \Delta) &= \frac{d^n}{d\lambda^n}(-g'(\lambda)R^2(g(\lambda), \Delta)) \\ &= \sum_{k=0}^{n} \sum_{h=0}^{n-k} a_{k,h}(-g^{(k+1)}(\lambda)) \frac{d^h}{d\lambda^h} R(g(\lambda), \Delta) \frac{d^{n-h-k}}{d\lambda^{n-h-k}} R(g(\lambda), \Delta) \end{aligned} \tag{3.24}$$

where the coefficients $a_{k,h}$ are positive integers. Since $-g^{(k+1)}(\lambda)\geqq 0$ for k even, and $-g^{(k+1)}(\lambda)\leqq 0$ for k odd, then (3.24) and the induction assumption also imply (3.21) for $k=n+1$. This completes the proof. □

Propositions 3.1 and 3.2 imply, together with Propositions 1.1, 1.2, and 2.1, the following results.

PROPOSITION 3.3. *Let* (3.8), (3.9), (3.15), *and* (3.4) *hold, and let* Ω *be a bounded open set in* $\mathbb{R}^n$ *with boundary* $\partial\Omega$ *of class* C^2. *Let* $f:\mathbb{R}=\bar{\Omega}\to\mathbb{R}$ *be continuous, bounded, and such that*

$$\sup_{s<t,x\in\bar{\Omega}} |f(t,x)-f(s,x)|(t-s)^{-\alpha}<+\infty.$$

Then

(a) *Problem* (3.6) *has a unique bounded solution* $u:\mathbb{R}\times\bar{\Omega}\to\mathbb{R}$. *Moreover,* u_t *and* Δu *are bounded in* $\mathbb{R}\times\bar{\Omega}$, *and*

$$\sup_{s<t,x\in\bar{\Omega}} |u_t(t,x)-u_t(s,x)|(t-s)^{-\alpha}+\sup_{s<t,x\in\bar{\Omega}} |\Delta u(t,x)-\Delta u(s,x)|(t-s)^{-\alpha}<+\infty.$$

(b) *If* f *is T-periodic with respect to* t, *then* u *is T-periodic with respect to* t; *if* $f(t,x)$ *converges to* $f_\infty(x)$ *as* $t\to+\infty$ (*respectively, to* $f_{-\infty}(x)$ *as* $t\to-\infty$), *uniformly for* $x\in\bar{\Omega}$, *then* $u(t,x)$ *converges to*

$$\bar{u}(x)=b_0\left(c_0-\int_0^{+\infty}\gamma(s)\,ds\right)^{-1}g(x),\qquad x\in\bar{\Omega}$$

(*uniformly for* $x\in\bar{\Omega}$), *as* $t\to+\infty$ (*respectively,* $t\to-\infty$), *where* g *is the solution of* $\Delta g=-f_\infty$ (*respectively,* $\Delta g=-f_{-\infty}$) *in* Ω, $g_{|\partial\Omega}=0$. *Moreover*

$$\lim_{t\to+\infty}\Delta u(t,\cdot)=-b_0\left(c_0-\int_0^{+\infty}\gamma(s)\,ds\right)^{-1}f_\infty,\qquad \lim_{t\to+\infty}u_t(t,\cdot)=0\quad \text{in } C(\bar{\Omega}),$$

respectively,

$$\lim_{t\to-\infty}\Delta u(t,\cdot)=-b_0\left(c_0-\int_0^{+\infty}\gamma(s)\,ds\right)^{-1}f_{-\infty},\qquad \lim_{t\to-\infty}u_t(t,\cdot)=0\quad \text{in } C(\bar{\Omega}).$$

(c) *If* $f(t,x)\geqq 0$ *for all* $t\in\mathbb{R}$, $x\in\bar{\Omega}$, *and* (3.19) *holds, then* $u(t,x)\geqq 0$ *for all* $t\in\mathbb{R}$, $x\in\bar{\Omega}$.

PROPOSITION 3.4. *Let* (3.8), (3.9), (3.15), *and* (3.4) *hold, and let* $\Omega\subset\mathbb{R}^n$ *be a bounded open set with* C^2 *boundary* $\partial\Omega$. *Let* $u_0\in C(\bar{\Omega})$ *be such that* $u_{0|\partial\Omega}=0$. *Let* $h:[0,+\infty[\times\bar{\Omega}\to\mathbb{R}$ *be continuous and such that* $t\to h(t,x)$ *is locally* α*-Hölder continuous* (*uniformly with respect to* $x\in\bar{\Omega}$) *for some* $\alpha\in\,]0,1[$. *Then*

(a) *Problem* (3.7) *has a unique solution* $u\in C([0,+\infty[\times\bar{\Omega})$, *such that* $t\to u_t(t,x)$ *and* $t\to\Delta u(t,x)$ *are locally* α*-Hölder continuous in* $]0,+\infty[$, *uniformly with respect to* $x\in\bar{\Omega}$.

(b) *If* $\lim_{t\to+\infty}h(t,\cdot)=h_\infty$ *in* $C(\bar{\Omega})$, *then, denoting by* g *the solution of* $\Delta g=-h_\infty$ *in* Ω, $g_{|\partial\Omega}=0$, *we have*

$$\lim_{t\to+\infty}u(t,\cdot)=b_0\left(c_0-\int_0^{+\infty}\gamma(s)\,ds\right)^{-1}g\qquad \text{in } C(\bar{\Omega}),$$

$$\lim_{t\to+\infty}\Delta u(t,\cdot)=-b_0\left(c_0-\int_0^{+\infty}\gamma(s)\,ds\right)^{-1}h_\infty\qquad \text{in } C(\bar{\Omega}),$$

$$\lim_{t\to+\infty}u_t(t,\cdot)=0\qquad \text{in } C(\bar{\Omega}).$$

(c) *If* (3.19) *holds,* $u_0(x) \geqq 0$ *for all* $x \in \bar{\Omega}$, *and* $h(t, x) \geqq 0$ *for all* $t \geqq 0$, $x \in \bar{\Omega}$, *then* $u(t, x) \geqq 0$ *for all* $t \geqq 0$, $x \in \bar{\Omega}$.

Remark 3.5. We need not assume complete monotonicity on β and γ to get the results of the previous propositions. Reading the proofs, we can see that it can be replaced by the following:

(3.25)
(a) β and γ are positive, nonincreasing, and log convex.
(b) The Laplace transforms of β and γ are analytically extendible to a sector $S = \{\lambda \in \mathbb{C};\ \lambda \neq 0,\ |\arg \lambda| < \theta\}$ with $\theta > \pi/2$, and the extensions (denoted by $\hat{\beta}(\lambda)$ and $\hat{\gamma}(\lambda)$, respectively) satisfy

$$\sup_{\lambda \in s} |\lambda|^{\alpha} |\hat{\beta}(\lambda)| < +\infty, \qquad \sup_{\lambda \in S} |\lambda|^{\delta} |\hat{\gamma}(\lambda)| < +\infty$$

for some α, $\delta > 0$.

On the other hand, (3.25b) implies that β and γ are analytic functions having a holomorphic extension to some sector in the complex plane around the positive real semi-axis.

Acknowledgments. The author expresses her thanks to Professor Ph. Clément for his encouragement and valuable remarks.

REFERENCES

[1] PH. CLÉMENT, *On abstract Volterra equations with kernels having a positive resolvent*, Israel J. Math., 36 (1980), pp.193-200.
[2] PH. CLÉMENT, R. C. MACCAMY, AND J. A. NOHEL, *Asymptotic property of solutions of nonlinear abstract Volterra equations*, J. Integral Equations, 3 (1981), pp. 185-216.
[3] PH. CLÉMENT AND J. A. NOHEL, *Abstract linear and nonlinear Volterra equations preserving positivity*, SIAM J. Math. Anal., 10 (1979), pp. 365-388.
[4] ———, *Asymptotic behavior of solutions of nonlinear Volterra equations with completely positive kernels*, SIAM J. Math. Anal., 12 (1981), pp. 514-535.
[5] B. D. COLEMAN, *Thermodynamics of materials with memory*, Arch. Rational Mech. Anal., 17 (1964), pp. 1-46.
[6] B. D. COLEMAN AND M. E. GURTIN, *Equipresence and constitutive equations for rigid heat conductors*, Z. Angew. Math. Phys., 18 (1967), pp. 199-207.
[7] B. D. COLEMAN AND V. J. MIZEL, *Thermodynamics and departures from Fourier's law of heat conduction*, Arch. Rational Mech. Anal., 13 (1963), pp. 245-261.
[8] J. F. D. DUFF, *Positive elementary solutions and completely monotonic functions*, J. Math. Anal. Appl., 27 (1969), pp. 469-494.
[9] G. DA PRATO AND M. IANNELLI, *Existence and regularity for a class of integrodifferential equations of parabolic type*, J. Math. Anal. Appl., 112 (1985), pp. 36-55.
[10] G. DA PRATO AND A. LUNARDI, *Solvability on the real line of a class of linear Volterra integrodifferential equations of parabolic type*, Ann. Mat. Pura Appl. (4), 150 (1988), pp. 67-117.
[11] M. E. GURTIN AND A. C. PIPKIN, *A general theory of heat conduction with finite wave speeds*, Arch. Rational Mech. Anal., 31 (1968), pp. 113-126.
[12] A. LUNARDI, *Laplace transform methods in integrodifferential equations*, J. Integral Equations, 10 (1985), pp. 185-211.
[13] S. O. LONDEN AND J. A. NOHEL, *A nonlinear Volterra integrodifferential equation occurring in heat flow*, J. Integral Equations, 6 (1984), pp. 11-50.
[14] R. C. MACCAMY, *An integro-differential equation with applications in heat flow*, Quart. Appl. Math., 35 (1977), pp. 1-19.
[15] J. W. NUNZIATO, *On heat conduction in materials with memory*, Quart. Appl. Math., 29 (1971), pp. 187-304.
[16] E. SINESTRARI, *On the abstract Cauchy problem of parabolic type in spaces of continuous functions*, J. Math. Anal. Appl., 107 (1985), pp. 16-66.
[17] H. B. STEWART, *Generation of analytic semigroups by strongly elliptic operators*, Trans. Amer. Math. Soc., 199 (1974), pp. 141-162.
[18] D. V. WIDDER, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1946.

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1225-1234, September 1990
© 1990 Society for Industrial and Applied Mathematics
007

SYSTEMS OF DIFFERENTIAL EQUATIONS THAT ARE COMPETITIVE OR COOPERATIVE. IV: STRUCTURAL STABILITY IN THREE-DIMENSIONAL SYSTEMS*

MORRIS W. HIRSCH†

Abstract. It is shown that among three-dimensional systems that are competitive or cooperative, those satisfying the generic Kupka-Smale conditions also satisfy the Morse-Smale conditions and are therefore structurally stable. This identifies a new and easily recognizable class of systems approximable by structurally stable systems.

Key words. stability, competitive, cooperative, dynamical system

AMS(MOS) subject classifications. 34, 58, 92

Introduction. A century ago Poincaré initiated what he called the qualitative theory of differential equations. He pointed out that very few systems of equations have solutions expressible in closed form, or as power series with recursively computable coefficients. Our only recourse, he suggested, is to understand the topological nature of the family of trajectories. To this end Poincaré developed many important tools for the theoretical analysis of orbit structure: hyperbolic periodic orbits, transverse stable and unstable manifolds, limit points, recurrence, and stability. He informally introduced the fruitful idea of ignoring "infinitely improbable" situations. These ideas were extended and exploited by G. D. Birkhoff. During the explosive modern revival of differentiable dynamics over the last three decades, many powerful new methods were introduced. But Poincaré's philosophical approach is still vital, and his technical concepts remain fundamental.

One of our advantages over Poincaré is that we have the benefit of a century's accumulation of dynamical lore: both general theory and detailed structure about many types of dynamical systems. As a result we can perceive, however dimly, that some of these systems fall into classes with similar dynamic behavior and similar phase portraits. As a result, for some systems (all too few!) we can successfully predict the dynamics from geometrical features of the phase portraits.

This outgrowth of Poincaré's approach has developed into an intensive research program into geometrical properties of phase portraits and their relation to dynamics. Poincaré's idea of ignoring improbable situations has become systematized into the search for useful "generic" properties. This is an inexact notion, but a generic property is often taken to be one that is shared by a Baire subset[1] of the space of vector fields under consideration; we use this definition for C^1 fields.

One of the important results of modern dynamics that we use below is the Kupka-Smale theorem: For C^r vector fields, $1 \leqq r \leqq \infty$, it is a generic property for all periodic orbits to be hyperbolic, with their stable and unstable manifolds meeting only transversely (Kupka (1963) and Smale (1963)). While for a given system it is rarely possible to verify these conditions, still the theorem is of great theoretical importance. Moreover, the fact that the conditions are generic—true for "most" systems—makes the Kupka-Smale theorem of practical importance as well: In the absence of reason

* Received by the editors May 12, 1989; accepted for publication December 1, 1989.

† Department of Mathematics, University of California, Berkeley, California 94720.

[1] A *Baire subset* of a complete metric space means the intersection of a countable family of dense open sets. By the Baire category theorem, such a subset is dense.

to assume the contrary, we might as well assume that the theorem holds for whatever system is under consideration.

Users of dynamics in many scientific disciplines attach great importance to *robust* dynamical properties: a feature of a dynamical system is robust if it persists under all sufficiently small perturbations of the system. The ultimate robust property is that of *structural stability*: A C^1 vector field X on a compact manifold M is *structurally stable* if it has a neighborhood $\mathcal{N}$ in the space of C^1 fields such that for every Y in $\mathcal{N}$ there is a homeomorphism of M taking oriented X-orbits to oriented Y-orbits.

Structurally stable dynamical systems were introduced by Andronov and Pontryagin (1937), who announced their density in the space of vector fields on the 2-disk which are transverse to the boundary; proof of this was published by De Baggis (1952). Peixoto (1962) proved that structurally stable systems are generic among C^1 vector fields on a compact orientable surface, and he gave simple necessary and sufficient conditions characterizing such systems: (1) All periodic orbits (cycles and equilibria) are hyperbolic, and (2) no orbits join saddle equilibria. These are in fact exactly the generic conditions of the Kupka–Smale theorem. In addition he showed that (3) the number of periodic orbits is finite, and (4) every limit set is a periodic orbit. A system satisfying conditions (1) through (4) is called *Morse–Smale*. Thus Peixoto proved that for systems on compact orientable surfaces, the structural stability, Kupka–Smale, and Morse–Smale concepts coincide. This was extended to nonorientable compact surfaces by Pugh (1967), (1967a) who repaired an error in Peixoto's proof for this case. An error in Pugh's proof was corrected by Pugh and Robinson (1983).

For a short time there was hope that structural stability could be proved generic, and characterized by simple properties. Unfortunately, it was soon shown that on manifolds of dimension greater than 2, structural stable systems are not dense (Newhouse (1970), (1980)). Moreover, even structurally stable systems can have very complex dynamics: The celebrated Anosov flows are structurally stable (Anosov (1967)), but the number of cycles in such a flow is infinite: there are limit sets that are not cycles or equilibria, and it can happen that almost every point has a dense orbit.

By now there are several general theoretical characterizations of structural stability and related properties for various kinds of systems on higher dimensional manifolds: see Palis (1970), Palis and Smale (1968), Smale (1970), Franks (1971), Robbin (1971), (1972), Pliss (1972), (1972a), Robinson (1971), (1976), Shub (1978), Liao (1980), Mañe (1988). But the dynamical hypotheses in these results are generally exceedingly hard to verify for any given vector field; and unlike the Kupka–Smale hypotheses, they are not generic. Moreover, very few specific examples of structurally stable systems are known, apart from special systems such as certain geodesic flows.

In Part I of this series (Hirsch (1982)) I distinguished between two methodologies in dynamics: the structural approach discussed above, which emphasizes structural stability and theoretical geometrical properties such as hyperbolicity and transversality; and the algebraic techniques that are common in applied dynamics. The structural approach has led to deep conceptual analyses of the structure of many classes of systems; but it is often of little use in analyzing a given system because of the extreme difficulty in verifying the requisite hypotheses. The algebraic techniques have proved useful in analyzing the orbit structure of many different models in biology, chemistry and other fields; but as these methods are often ad hoc, few general principles have emerged.

In this series of articles (Hirsch (1982), (1985), (1988a), (1989); see also (1984), (1988)) I have combined both approaches to analyze a class of systems that are often

used as models in applied fields, namely cooperative and competitive systems (defined below). These systems are characterized by simple inequalities on the partial derivatives of their vector fields.

We now come to the purpose of this article: to show that *there is a certain easily recognizable class $\mathscr{C}$ of C^1 vector fields on three-dimensional Euclidean space $\mathbb{R}^3$—namely, dissipative competitive systems—in which systems that are structurally stable are generic.* In fact the Kupka-Smale fields in $\mathscr{C}$ are precisely the structurally stable ones. Moreover, these fields have the property that every limit set is a closed orbit, and these are finite in number; in other words, these fields are not only Kupka-Smale but Morse-Smale. Thus $\mathscr{C}$ satisfies an exact analogue of Peixoto's theorem.

Parts I-III of this series analyzed the geometry and dynamics of compact limit sets of cooperative and competitive vector fields by exploiting the order-preserving properties of their flows. In particular, it was shown in Part I that such a limit set K projects homeomorphically into a hyperplane transverse to any positive vector; moreover, the dynamics in K are isomorphic to the dynamics in the image K' of K under a Lipschitz vector field in the hyperplane, having K' as an invariant set. In this article I apply this result to the flow defined by a cooperative or competitive vector field in $\mathbb{R}^3$, taking advantage of the simplicity of planar dynamics to study limit sets. For most of the results I do not need the hypothesis of irreducibility which was used heavily in Parts II and III.

The basic result of this paper is Theorem 1: Under generic conditions all limit sets are equilibria or cycles, and the number of cycles in any compact set is finite. (A similar result is due to Smith and Waltman (1987).) In Part V (Hirsch (1989)) further implications will be developed, and structural stability of certain feedback loops will be proved.

We apply Theorem 1 to the case where the vector field is transverse to the boundary of a compact 3-manifold $M \subset \mathbb{R}^3$. Theorem 2 asserts that if the flow in M satisfies the generic conditions of Kupka-Smale, then it satisfies the stronger conditions of Morse-Smale. Such a flow is therefore structurally stable, and the only limit sets are a finite set of cycles and equilibria. Thus for competitive or cooperative fields transverse to ∂M, the structural stability, Morse-Smale, and Kupka-Smale concepts coincide.

Since Kupka-Smale fields are generic, it follows that if F is any competitive or cooperative field transverse to ∂M, then $F|M$ is in the closure of the Morse-Smale fields (for the C^1 topology). Theorem 3 shows that if in addition F is irreducible, then $F|M$ is in the interior of the closure of the Morse-Smale fields. This means that not only F, but also any field sufficiently C^1 close to F, can be approximated by structurally stable fields having only a finite number of periodic orbits.

Theorem 4 shows that the Kupka-Smale conditions also characterize competitive and cooperative fields that are structurally stable, in a certain sense, in noncompact domains in $\mathbb{R}^3$. Theorem 5 shows that this holds in the basin of any attractor.

We use the following notation for the vector order in $\mathbb{R}^3$:

$$x \leqq y \quad \text{if } x_i \leqq y_i \quad (i = 1, 2, 3),$$

$$x < y \quad \text{if } x \leqq y \quad \text{and} \quad x \neq y,$$

$$x < y \quad \text{if } x_i < y_i \quad (i = 1, 2, 3).$$

Notations such as $y > x$ have the natural meanings.

The *closed nonnegative cone* is $\mathbb{R}^3_+ = \{x \in \mathbb{R}^3 : x \geqq 0\}$; setting $x > 0$ defines its interior.

A set C is *p-convex* if it contains the line segment with endpoints a, b whenever $a \in C$, $b \in C$, and $a < b$.

For any points u, v in $\mathbb{R}^3$ define the

$$\textit{closed order interval } [u, v] = \{x: u \leqq x \leqq v\}$$

and the

$$\textit{open order interval } [[u, v]] = \{x: u < x < v\}.$$

A set $K \subset \mathbb{R}^3$ is *balanced* if no two points of K are related by $<$, and *strongly balanced* if no two points are related by $<$.

Throughout this paper $X \subset \mathbb{R}^3$ is a nonenpty p-convex open set.

We denote by $F: X \to \mathbb{R}^3$ a C^1 vector field generating the (local) flow $\varphi = \{\varphi_t\}$ in X. Thus the solution to the initial value problem $\dot{u} = F(u)$, $u(0) = x$ is the curve $t \mapsto \varphi_t x$, defined for t in some open interval $I_x = (\sigma_x, \tau_x)$, $-\infty \leqq \sigma_x < 0 < \tau_x \leqq \infty$.

We call F (or φ) *cooperative* if

$$\partial F_i / \partial x_j \geqq 0 \quad \text{for } i \neq j.$$

Since X is p-convex, the Müller-Kamke theorem implies that a cooperative flow φ is *monotone*, i.e., φ_t for $t \geqq 0$ preserves the vector order of $\mathbb{R}^3$ (Müller (1926), Kamke (1932); see also Coppel (1965)). In fact, φ has the property that for $t \geqq 0$ the Jacobian matrices $D\varphi_t(x)$ have nonnegative entries, denoted by $D\varphi_t(x) \geqq 0$. In a p-convex domain this implies monotonicity of φ; see Part II.

For some results we need the additional assumption that F is *irreducible*, i.e., the Jacobian matrices $DF(x)$ are irreducible. In Part II it is shown that when this holds and F is cooperative then $D\varphi_t(x) > 0$ for $t > 0$, and p-convexity then implies that φ is *strongly monotone*:

$$\varphi_t x < \varphi_t y \quad \text{if } x < y \quad \text{and} \quad t > 0.$$

A vector field H is *competitive* if $-H$ is cooperative, that is, if $\partial H_i / \partial x_j \leqq 0$ for $i \neq j$. Many propositions about cooperative fields are invariant under time reversal (replacing the field by its negative) and thus are also true for competitive fields.

A *circuit* is a finite sequence of equilibria $p_0, \cdots, p_n = p_0$, $n \geqq 1$, such that $W^u(p_{i-1}) \cap W^s(p_i)$ is not empty, where W^u and W^s denote stable and unstable manifolds. (In Part I this was called a "cycle of equilibria"; in this paper, the term "cycle" refers to a nontrivial periodic orbit.) If all equilibria are hyperbolic and their stable and unstable manifolds mutually transverse (which is a generic condition; see Kupka (1963), Smale (1963), or Abraham and Robbin (1967)), then there cannot be any circuits.

From now on, F denotes a C^1 vector field in X which is competitive or cooperative; the flow generated by F is denoted by φ.

We begin with the characteristic geometry of limit sets:

PROPOSITION 0. *For the system F:*

(a) *Every cycle is strongly balanced and every compact limit set is balanced.*

(b) *If the system is irreducible then every compact limit set is strongly balanced.*

Proof. It suffices to consider a cooperative field F. For cooperative systems the two statements about limit sets follow from Theorem 2.3(b)(c) of Part II. To see that a cycle K in a cooperative system is strongly balanced, let $\lambda > 0$ be the minimum period of K. If $z \in K$, $0 \leqq t < \lambda$ and $\varphi_t z = z$, then $t = 0$. Suppose $x \leqq y$ in K. Then we can find $0 \leqq t < \lambda$ such that $\varphi_t x = y$. As $n \to \infty$ through positive integers, the monotone sequence $\varphi_{tn} x$ converges to a point $z \in K$ fixed under φ_t. But then $t = 0$; thus $x = y$. □

Next we complete a train of thought begun in Part I. Let K be a compact limit set of a cooperative system, with no equilibrium in K. Then by Theorem C of Part I, K is either a cycle or a cylinder of cycles; moreover, if K is an ω-limit set then it must be a cycle. Further conditions (on the linearized Poincaré map) under which K must be a cycle can be derived from Hale and Stokes (1960), as was pointed out to me by H. L. Smith. However, Smith (1986) showed that when K is an α-limit set and the vector field is irreducible, then K must be a cycle; and the following theorem sharpens Smith's result by eliminating the assumption of irreducibility. We thus have a general result valid for both cooperative and competitive systems.

THEOREM 1. *A compact limit set K containing no equilibrium is a cycle.*

Proof. We assume F is cooperative, otherwise replacing it by $-F$.

According to Theorem C of Part I, if K is not a cycle then K is an α-limit set consisting of a cylinder of cycles. We assume K has these properties and derive a contradiction.

Let v be a positive vector, E the plane through the origin perpendicular to v, and $\pi:\mathbb{R}^3\to E$ the orthogonal projection. Then π maps K homeomorphically onto a compact subset of E homeomorphic to an annulus (Part I, Theorem A). Let $C\subset K$ be a cycle mapping under π into the interior of πK. The Jordan curve πC separates πK. Fix points a and b in $K\setminus C$ such that πa and πb are in different components of $\pi K\setminus \pi C$.

Since as $t\to-\infty$ the negative semiorbit $\varphi_{-t}x$ of x repeatedly visits every neighborhood of a and b, it follows that $\pi\varphi_{-t}x$ crosses πC at a sequence of times $-t_k\to-\infty$. This means that there are points $z_k\in C$ related to $\varphi_{-t_k}x$ by $<$ or $>$, for all k. Infinitely many of these relations are the same, $<$ or $>$. Passing to a subsequence we assume they are all $>$, the other case being similar. Thus we assume that $\varphi_{-t_k}x<z_k$.

For every $s>0$ there is a point $w\in C$ such that $\varphi_{-s}x>w$. To see this, choose k so large that $t_k>s$. Then

$$\varphi_{-s}x=\varphi_{t_k-s}\varphi_{-t_k}x<\varphi_{t_k-s}z_k\in C$$

by monotonicity of φ_{t_k-s} and invariance of C, proving the assertion.

Since $K=\alpha(x)$, *it follows that every point of K is less than or equal to some point of C.*

The same reasoning applies to every cycle $C'\subset K$: Either every point of K is less than or equal to some point of C' or every point of K is greater than or equal to some point of C'. Since we can find three different cycles in K, there are two of them for which this holds with the same relation $\leqq$ or $\geqq$. We consider the case of $\leqq$, the other being similar.

We assume, then, that there are distinct cycles C_1, C_2 in K such that every point of K is less than or equal to some point of C_1 and less than or equal to some point of C_2. For any $u\in C_1$ we can therefore find $w\in C_2$ and $z\in C_1$ such that $u<w<z$. But $u<z$ is impossible because every cycle is strongly balanced, by Proposition 0. This contradiction shows that K must consist of a single cycle. □

THEOREM 2. *Let $\Gamma\subset X$ be a compact set. Assume*:

(i) *All equilibria in Γ are hyperbolic and there are no circuits*;

(ii) *For any real number $T>0$ the number of cycles in Γ having period less than or equal to T is finite.*

Then:

(a) *Every limit set in Γ is an equilibrium or cycle*;

(b) *The number of cycles in Γ is finite.*

Note that conditions (i) and (ii) are generic; condition (ii) holds if all cycles are hyperbolic, or more generally, if every cycle has a linearized Poincaré map for which 1 is not an eigenvalue.

Proof. We assume F is cooperative, replacing it with $-F$ otherwise.

We first prove (b). Suppose that in Γ there is an infinite sequence of distinct cycles. Then we can choose such a sequence $\{C_n\}$ together with points $x_n \in C_n$ converging to a point $y \in \Gamma$. Let $P \subset \Gamma$ denote the closure of $\cup C_n$.

LEMMA. *For any n and any $x \in C_n$ there exist $m > n$ and $y \in C_m$ such that $x > y$ or $x < y$.*

To prove the lemma, suppose the contrary. Passing to a subsequence we can assume that no two points of P are related by $<$. By Theorem A of Part I, there is a Lipschitz vector field G in the plane, with flow ψ, and a homeomorphism g from P onto a compact set P' in the plane, such that P' is invariant under ψ, and $\psi_t g(x) = g\varphi_t(x)$ for all $x \in P$. It follows that P' is the closure of the distinct cycles $g(C_n) = C_n'$ for ψ. Set $x_n' = g(x_n)$, so that $x_n' \to y' = g(y)$. Now the orbit of y' cannot be a cycle C', for if it were, its period would be the limit of the periods of the C_n'. But this would imply that the period of the cycle $g^{-1}C'$ is the limit of the periods of the C_n, contrary to hypothesis (ii).

Denote by L the set of points obtainable as limits of sequences $v_n \in C_n$; it is easy to see that L is a nonempty, compact invariant set. Set $L' = g(L)$. The argument just given shows that L' contains no cycles. Moreover, for any $u \in L'$ it must be that $\alpha(u)$ and $\omega(u)$ are equilibria: Otherwise, by the generalized Poincaré-Bendixson theorem (Hartman (1964)), L' would contain a circuit, contradicting the hypothesis that Γ contains no circuit (see the remark following the proof). It follows that the alpha and omega limit sets of every point of L are equilibria.

It is easy to see that if $p \in L$ is an equilibrium, then L must contain points of $W^s(p)\backslash p$ and points of $W^u(p)\backslash p$, where W^s and W^u denote stable and unstable manifolds.

We show next that L contains a circuit. Let $p_0 \in L$ be an equilibrium. Choose $x_0 \in W^u(p_0)\backslash p_0$ and set $p_1 = \omega(x_0)$. Then $p_1 \in L$ is an equilibrium and $W^u(p_0) \cap W^s(p_1) \neq \varnothing$. Repeating this construction with p_1 in place of p_0, we obtain an equilibrium $p_2 \in L$ with $W^u(p_1) \cap W^s(p_2) \neq \varnothing$. Since the equilibrium set is finite (by hyperbolicity and compactness) we arrive at a circuit, contrary to hypothesis. This proves the lemma.

Continuing the proof of part (b) of Theorem 1, by the preceding lemma we can pass to a subsequence and assume that there are points $x_n \in C_n$, $w_{n+1} \in C_{n+1}$ with $x_n < w_{n+1}$ for all n, or $x_n > w_{n+1}$ for all n. The two cases are similar; we assume the former.

We can assume $w_{n+1} = x_{n+1}$ for all n: Recursively assume this holds for $n = 1, \cdots, k-1$. Now $x_k < w_{k+1}$, and w_{k+1} is on the same cycle as x_{k+1}. Therefore there is a positive number s such that $\varphi_s x_{k+1} = w_{k+1}$. Since φ_s preserves order, $w_{k+1} < \varphi_s w_{k+2}$. We complete the induction by replacing x_{k+1} with w_{k+1} and w_{k+2} with $\varphi_s w_{k+2}$.

By passing to a subsequence we assume $x_n \to q \in L$. We know that $\omega(q)$ is an equilibrium p. It follows that $C_n < p$ for all n: To see this, let $z \in C_n$ be arbitrary and let $t_i \to \infty$ be such that $\varphi_{t_i} x_n \to z$ as $i \to \infty$. Then $\varphi_{t_i} q \to p$, so $z \leqq p$ since $x_n < p$ and ϕ_{t_i} preserves $<$. For all n we now have $x_n < x_{n+1} \leqq p$. Since every point of C_n is on the forward orbit of x_n, it follows from monotonicity that $C_n < p$.

Now let $U \subset X$ be any cubical neighborhood of p. I claim $C_n \subset U$ for sufficiently large n. To see this let m be such that $x_m \in U$. For this m, let b be the supremum of C_m in the vector order, that is, each coordinate of b is the supremum of the corresponding coordinates of all the points of C_m. Then $b < p$. Also, since C_m is invariant, by monotonicity we have $\varphi_t b \geqq C_m$ for all $t \geqq 0$. Therefore $\varphi_t b \geqq b$ for all $t \geqq 0$, by the

definition of b. Fix n_0 so large that for $n \geqq n_0$ we have $b < x_n < p$. By monotonicity, $\varphi_t b < \varphi_t x_n < p$. Since $b \leqq \varphi_t b$, we have $b < \varphi_t x_n < p$. Therefore $b < C_n < p$ for all $n \geqq n_0$, and since U is cubical this implies $C_n \subset U$.

We have shown that every neighborhood of p contains a cycle. But this is impossible since p is a hyperbolic equilibrium. This contradiction completes the proof of (b).

To prove (a), let $K \subset \Gamma$ be a limit set. By Theorem A of Part I the dynamical system in K is topologically equivalent to the dynamics in a compact invariant set of the flow of a Lipschitz vector field in the plane. Therefore by the generalized Poincaré-Bendixson theorem (see remark below), either K is an equilibrium or a cycle or else K contains a circuit. Since the latter is contrary to hypothesis, the proof of Theorem 1 is complete. □

Remark. What Hartman proves (1964, Thm. 4.2) is the following: Let $K \subset \mathbb{R}^2$ be a nonempty compact limit set of a planar flow ψ, containing only a finite number of equilibria $p_1, \cdots, p_n$, $n \geqq 1$. Then $K \backslash \{p_1, \cdots, p_n\}$ consists of a countable number of orbits joining the p_j. To see that this implies K contains a circuit, observe that if p_j is the ω-limit of some orbit $\{\psi_t x_0\}$ in K, then p_j must be the α-limit of some orbit $\{\psi_t x_1\}$ in K; for otherwise p_j would be an attractor for the flow in K, which is impossible because K is a limit set. Thus starting from x_0, we recursively find in K a sequence of points x_k in K and equilibria q_k such that $\omega(x_k) = q_k = \alpha(x_{k+1})$. Since the set of q_k is finite we must obtain a circuit.

For any manifold Z let $\mathscr{V}^1(Z)$ denote the space of C^1 vector fields on Z which are transverse to ∂Z, where r is a positive integer, endowed with the weak C^1 topology (Hirsch (1976)).

Let $M \subset X$ be a smooth compact three-dimensional submanifold with boundary ∂M.

Our next result characterizes cooperative and competitive vector fields $F \in \mathscr{V}^1(X)$ such that $F \mid M$ is *structurally stable*: for any $\varepsilon > 0$ there is a neighborhood $\mathscr{U}$ of $F \mid M$ in $\mathscr{V}^1(M)$ such that for any $H \in \mathscr{U}$ there is a homeomorphism h of M moving no point by more than ε, which takes every F-orbit in M onto an H orbit, preserving orientation.

A sufficient condition for structural stability is that a vector field on M satisfy the Morse-Smale conditions (see Palis (1969)):

MORSE-SMALE CONDITIONS.

(i) All equilibria and cycles are hyperbolic and their stable manifolds intersect only transversely;

(ii) Every limit set is an equilibrium or a cycle.

If only (i) is assumed then G is called *Kupka-Smale*, and these conditions hold for a Baire subset of $\mathscr{V}^1(M)$. (See Kupka (1963), Smale (1963); for another proof see Abraham and Robbin (1967).) The structural stability theorem for three-dimensional Morse-Smale systems, due to Palis (1969), is stated for manifolds without boundary; but it is well known that the proof adapts to manifolds with nonempty boundary, provided the vector field is transverse to the boundary. The theorem is true in all dimensions (Palis and Smale (1968)).

THEOREM 3. *Suppose that F is transverse to ∂M, and that $F \mid M$ is Kupka-Smale. Then $F \mid M$ is Morse-Smale, and therefore structurally stable. Conversely, if $F \mid M$ is structurally stable, then it is Morse-Smale.*

Proof. We assume F is cooperative, replacing it by $-F$ in the competitive case. Assume F is Kupka-Smale. Then the set of equilibria in M is finite, and their stable and unstable manifolds are mutually transverse. A well-known argument based on dimensions of stable and unstable manifolds therefore implies that there are no circuits

in M. By Theorem 2 (with $\Gamma = M$) the set of cycles in M is finite, and every limit set is a cycle or an equilibrium (Hypothesis ii of Theorem 2 holds by Theorem 24.2 of Abraham and Robbin (1967).) Therefore $F|M$ is Morse-Smale and so structurally stable.

Conversely, suppose $F|M$ is structurally stable. It follows by results of Markus (1961) that a structurally stable vector field is Kupka-Smale; and we have just seen that this implies Morse-Smale for a field that is cooperative or competitive. □

From this it follows that for any C^1 cooperative or competitive vector field G in X which is transverse to ∂M, $G|M$ is in the closure in $\mathcal{V}^1(M)$ of the structurally stable fields—in fact, of the Morse-Smale fields. A slightly stronger conclusion holds for irreducible fields.

THEOREM 4. *Let G be an irreducible C^1 cooperative or competitive vector field on X, transverse to ∂M. Then $G|M$ is in the interior of the closure in $\mathcal{V}^1(M)$ of the set of Morse-Smale vector fields on M. Thus any field in M sufficiently close to $G|M$ is a limit of fields that are structurally stable.*

Proof. It suffices to consider the cooperative case. In Theorem 1.2 of Part II, I showed that any field F in X sufficiently C^1 close to an irreducible cooperative field, in a given compact set $K \subset X$, has a flow φ with the following property: There exists $\varepsilon > 0$ such that if $t > \varepsilon$ and $\varphi_s(x) \in M$ for all $0 \leqq s \leqq t$, then the Jacobian matrix $D\varphi_t(x)$ has only positive entries. For K we take a compact p-convex neighborhood of the smallest p-convex set containing M; p-convexity then implies that $\varphi_t|K$ is strongly monotone for $t > \varepsilon$. For a flow φ with this property, the results on projections of limit sets used in the proof of Theorem 2 are valid; and the same proof of Theorem 3 goes through. □

The existence of the compact manifold M transverse to F is not as stringent an assumption as it might appear. It follows from Theorem 3.2 of Wilson (1969) that such a manifold exists as an arbitrarily small neighborhood of any *uniform attractor* Q: This term means that $Q \subset X$ is a nonempty compact invariant set having a neighborhood in X in which $\lim_{t\to\infty} \operatorname{dist}(\varphi_t(x), Q) = 0$ uniformly. The set of all points tending to Q is the *basin of attraction* of Q, denoted here by U. It is easy to see that then $U \setminus \operatorname{Int} M$ is diffeomorphic to $\partial M \times [0, \infty)$. Wilson proved that Q has a C^∞ Lyapunov function $h: U \to \mathbb{R}$: $h(\varphi_t x) < h(x)$ if $t > 0$, $x \in U \setminus Q$. By the theorem of Morse (1939) (extended by Sard (1942)), h has a regular value $\alpha \in \mathbb{R}$. Then $h^{-1}(\alpha)$ is the boundary of the compact 3-manifold $M = h^{-1}(-\infty, \alpha]$, which is a neighborhood of Q; and the flow enters M transversely along $\partial M = h^{-1}(\alpha)$.

If M and U are obtained in this way and $F|M$ is structurally stable, then we will show that the flow in U enjoys the following kind of structural stability.

DEFINITION. The flow $\varphi|Y$ (or the vector field $F|Y$) in an invariant open set $Y \subset X$ is *structurally stable* if for every $\varepsilon > 0$ and every compact subset $Q \subset Y$ there exists a neighborhood $\mathcal{N} \subset \mathcal{V}^1(Y)$ of $F|Y$ with the following properties: For any vector field $G \in \mathcal{N}$ there is a homeomorphism h of Y onto an open subset of Y such that $|h(x) - x| < \delta$ for $x \in Q$, and h maps orbits of F into orbits of G, preserving orientation of trajectories.

In the case where U is the basin of a uniform attractor, and M is as above with $F|M$ structurally stable, we obtain h as follows. Suppose we are given $\varepsilon > 0$ and $Q \subset U$ as in the preceding definition. Let $\mathcal{W} \subset \mathcal{V}^1(M)$ be a neighborhood of $F|M$ so small that $H|M$ is transverse to ∂M for any $H \in \mathcal{W}$, and by the structural stability of $F|M$ there exists a homeomorphism h of M within ε of the identity, taking trajectories of $F|M$ to trajectories of H. Suppose $G \in \mathcal{V}^1(Y)$ is so close to F that $G|M \in \mathcal{W}$. Set $H = G|M$ and let $h: M \to M$ be as above. Extend h to a map $g: V \to V$ as follows. We

assume F points out of M at points of ∂M, the other case being similar. For each point $x \in \partial M$, h maps the forward F-trajectory of x onto the forward G-trajectory of x, preserving length. It is easy to see that if $\mathcal{W}$ is a small enough neighborhood of $F|M$ then $h|Q$ is within ε of the identity, and h maps Y homeomorphically onto an open set in Y.

Thus we have proved the following result.

THEOREM 5. *Let F be a competitive or cooperative vector field in the open p-convex set $X \subset \mathbb{R}^3$. Let $U \subset X$ be the basin of a uniform attractor. Assume that $F|U$ satisfies the Kupka-Smale conditions. Then $F|U$ is structurally stable, there are only finitely many periodic orbits in U, and every limit set in U is an equilibrium or a cycle.*

REFERENCES

R. ABRAHAM AND J. ROBBIN (1967), *Transversal Mappings and Flows*, W. A. Benjamin, New York.

A. ANDRONOV AND L. PONTRYAGIN (1937), *Systèmes grossiers*, Dokl. Akad. Nauk. SSSR, 14, pp. 247-251.

D. V. ANOSOV (1967), *Geodesic flows on closed Riemannian manifolds with negative curvature*, Proc. Steklov Inst. Math., 90, pp. 1-235.

W. COPPEL (1965), *Stability and Asymptotic Behavior of Differential Equations*, Heath, Boston.

H. F. DE BAGGIS (1952), *Dynamical systems with stable structures*, in Contributions to the Theory of Nonlinear Oscillations, II, S. Lefschetz, ed., Princeton University Press, Princeton, NJ.

J. FRANKS (1971), *Necessary conditions for stability of diffeomorphisms*, Trans. Amer. Math. Soc., 158, pp. 301-308.

P. HARTMAN (1964), *Ordinary Differential Equations*, John Wiley, New York.

J. K. HALE AND A. P. STOKES (1960), *Behavior of solutions of differential equations near integral manifolds*, Arch. Rational Mech. Anal., 16, pp. 133-170.

M. W. HIRSCH (1976), *Differential Topology*, Springer-Verlag, Berlin, New York.

——— (1981), Technical Report PAM-16, Center for Pure and Applied Mathematics, University of California, Berkeley.

——— (1982), *Systems of differential equations that are competitive or cooperative.* I: *Limit sets*, SIAM J. Math. Anal., 13, pp. 167-179.

——— (1982a), *Convergence in ordinary and partial differential equations.* Lecture Notes for Colloquium Lectures at University of Toronto, August 23-26, 1982, American Mathematical Society, Providence, RI.

——— (1983), *Differential equations and convergence almost everywhere in strongly monotone semiflows*, Contemp. Math., 17, pp. 267-285.

——— (1984), *The dynamical systems approach to differential equations*, Bull. Amer. Math. Soc., 11, pp. 1-64.

——— (1985), *Systems of differential equations that are competitive or cooperative.* II: *Convergence almost everywhere*, SIAM J. Math. Anal., 16, pp. 423-439.

——— (1988), *Stability and convergence in strongly monotone dynamical systems*, J. Reine Angew. Math., 383, pp. 1-53.

——— (1988a), *Systems of differential equations that are competitive or cooperative.* III: *Competing species*, Nonlinearity, 1, pp. 51-71.

——— (1989), *Systems of differential equations that are competitive or cooperative.* V: *Convergence in 3-dimensional systems*, J. Differential Equations, 80, pp. 94-106.

E. KAMKE (1932), *Zur Theorie der Systeme gewöhnlicher differential Gleichungen* II. Acta Math., 58, pp. 57-85.

I. KUPKA (1963), *Contributions à la théorie des champs génériques*, Contrib. Differential Equations, 2, pp. 457-484.

S. T. LIAO (1980), *On the stability conjecture*, Chinese Ann. Math., 1, pp. 9-30.

R. MAÑÉ (1988), *A proof of the C^1 stability conjecture*, Publ. Math. Inst. Hautes Etudes Scientifiques, 66, pp. 161-210.

L. MARKUS (1961), *Structurally stable dynamical systems*, Ann. of Math., 73, pp. 1-19.

A. P. MORSE (1939), *The behavior of a function on its critical set*, Ann. of Math., 40, pp. 62-70.

M. MÜLLER (1926), *Über das Fundamentaltheorem in der Theorie der gewöhnlichen Differentialgleichungen*, Math. Z., 26, pp. 619-645.

S. NEWHOUSE (1970), *Nondensity of Axiom $A(a)$ on S^2*, in Proc. Sympos. Pure Math., 14, Global Analysis, S.-S. Chern and S. Smale, eds., American Mathematical Society, Providence, RI.

——— (1980), *Lectures on dynamical systems*, Progress in Math., 8, pp. 1-114.

J. PALIS (1969), *On Morse-Smale dynamical systems*, Topology, 8, pp. 385-405.

——— (1970), *A note on Ω-stability*, in Proc. Sympos. Pure Math., 14, Global Analysis, S.-S. Chern and S. Smale, eds., American Mathematical Society, Providence, RI.

J. PALIS AND S. SMALE (1968), *Structural stability theorems*, in Proc. Sympos. Pure Math., 14, Global Analysis, S.-S. Chern and S. Smale, eds., American Mathematical Society, Providence, RI.

M. M. PEIXOTO (1962), *Structural stability on two-dimensional manifolds*, Topology, 1, pp. 101-120.

V. A. PLISS (1972), *Analysis of the necessity of the conditions of Smale and Robbin for structural stability of periodic systems of differential equations*, Differentsial'nye Uravneniya, 8, pp. 972-983.

——— (1972a), *On a conjecture due to Smale*, Differentsial'nye Uravneniya, 8, pp. 268-282.

C. PUGH (1967), *The closing lemma*, Amer. J. Math., pp. 956-1009.

——— (1967a), *Structural stability on M^2*, Anais Academia Brasileira Ciências, 39, pp. 45-48.

C. PUGH AND C. ROBINSON (1983), *The C^1 closing lemma including Hamiltonians*, Ergodic Theory Dynamical Systems, 3, pp. 261-313.

J. ROBBIN (1971), *A structural stability theorem*, Ann. of Math., 94, pp. 447-493.

——— (1972), *Topological conjugacy and structural stability*, Bull. Amer. Math. Soc., 78, pp. 923-952.

C. ROBINSON (1971), *C^r structural stability implies Kupka-Smale*, in Dynamical Systems, Salvador, Academic Press, New York.

——— (1976), *Structural stability of C^1 diffeomorphisms*, J. Differential Equations, 22, pp. 28-73.

A. SARD (1942), *The measure of the critical points of differentiable maps*, Bull. Amer. Math. Soc., 48, pp. 883-890.

J. SELGRADE (1980), *Asymptotic behavior of solutions to single loop positive feedback systems*, J. Differential Equations, 38, pp. 80-103.

M. SHUB (1978), *Stabilité globale des systèmes dynamiques*, Astérisque, 56.

S. SMALE (1963), *Stable manifolds for differential equations and diffeomorphisms*, Ann. Scuole Norm. Pisa (3), 17, pp. 97-116.

——— (1970), *The Ω-stability theorem*, in Proc. Sympos. Pure Math., 14, Global Analysis, S.-S. Chern and S. Smale, eds., American Mathematical Society, Providence, RI.

H. L. SMITH (1986), *Periodic orbits of competitive and cooperative systems*, J. Differential Equations, 65, pp. 361-373.

——— (1986a), *On the asymptotic behavior of a class of deterministic models of cooperating species*, SIAM J. Appl. Math., 46, pp. 368-375.

——— (1986b), *Periodic solutions of periodic competitive and cooperative systems*, SIAM J. Math. Anal., 17, pp. 1289-1318.

——— (1986c), *Periodic competitive differential equations and the discrete dynamics of competitive maps*, J. Differential Equations, 64, pp. 165-194.

——— (1986d), *Competing subcommunities of mutualists and a generalized Kamke Theorem*, SIAM J. Appl. Math., 46, pp. 856-874.

——— (1988), *Systems of differential equations which generate a monotone flow: A survey of results*, SIAM Rev., 30, pp. 87-113.

H. L. SMITH AND P. WALTMAN (1987), *A classification theorem for three-dimensional competitive systems*, J. Differential Equations, 70, pp. 325-332.

W. WILSON (1969), *Smoothing derivatives of functions and applications*, Trans. Amer. Math. Soc., 139, pp. 413-428.

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1235-1244, September 1990
© 1990 Society for Industrial and Applied Mathematics
008

LIMIT CYCLES FOR A CLASS OF ABEL EQUATIONS*

A. GASULL† AND J. LLIBRE†

Abstract. The number of solutions of the Abel differential equation $dx(t)/dt = A(t)x(t)^3 + B(t)x(t)^2 + C(t)x(t)$ satisfying the condition $x(0) = x(1)$ is studied, under the hypothesis that either $A(t)$ or $B(t)$ does not change sign for $t \in [0, 1]$. The main result obtained is that there are either infinitely many or at most three such solutions. This result is also applied to control the maximum number of limit cycles for some planar polynomial vector fields with homogeneous nonlinearities.

Key words. Abel differential equation, limit cycle, Riccati equation

AMS(MOS) subject classifications. primary 34C05, secondary 58F21

1. Introduction and statement of the main results.

A problem proposed by Pugh (see [12]) consists of the following: Let $a_0, a_1, \cdots, a_n : \mathbf{R} \to \mathbf{R}$ be smooth functions and consider the differential equation

$$\frac{dx}{dt} = a_n(t)x^n + a_{n-1}(t)x^{n-1} + \cdots + a_1(t)x + a_0(t), \qquad 0 \leqq t \leqq 1. \tag{1}$$

We will say that a solution $x(t)$ of (1) is a *closed solution* or a *periodic solution* if it is defined in the interval $[0, 1]$ and $x(0) = x(1)$. The adjectives "closed" and "periodic" are motivated by the case where $a_0, a_1, \cdots, a_n$ are 1-periodic, in which (1) can be considered in the cylinder and the "closed" solutions really correspond to periodic orbits in the cylinder. An isolated closed solution in the set of all the closed solutions will be called a *limit cycle.* Then the problem is: Does there exist a bound on the number of limit cycles of (1)?

In the case $n = 2$, (1) is called the *Riccati equation* and the problem of determining the number of limit cycles is already known: there are at most two of them (see, for instance, [12], [14]). When $n = 3$, (1) is called the *Abel equation.* Also in [12] it is proved that there is no upper bound for the number of closed solutions for the Abel equations. Hence a more specific problem arises: Give a bound on the number of limit cycles of Abel equations assuming additional hypotheses on $a_3(t)$, $a_2(t)$, $a_1(t)$, and $a_0(t)$.

A problem that is studied in several papers is Pugh's problem for Abel equations when $a_3(t)$ does not change sign (see [7], [12], [18]). In this case the maximum number of closed solutions is three.

The Ricatti equation acquired importance when it was introduced by Jacopo Francesco, Count Riccati of Venice (1676-1754), who worked in acoustics, to help solve second-order ordinary differential equations. Abel's differential equation arose in the context of the studies of N. H. Abel on the theory of elliptic functions.

The aim of this paper is to study the problem of determining the maximum number of limit cycles of Abel equations when $a_0(t) \equiv 0$ and one of the other three functions that define the differential equation does not change sign. For simplicity we write the

* Received by the editors October 31, 1988; accepted for publication (in revised form) October 30, 1989. The work of the two authors was partially supported by Dirección General de Investigación Científica y Tecnológica (DGICYT) grant PB86-0351.

† Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain.

Abel equation with $a_0(t) \equiv 0$ in the following form:

$$\frac{dx}{dt} = A(t)x^3 + B(t)x^2 + C(t)x. \tag{2}$$

Note that any Abel equation with a periodic orbit $x_1(t)$ can be written in the form (2) by using the new coordinate $\bar{x} = x - x_1(t)$. Observe also that the function $A(t)$ does not change in the new coordinate.

Let L and L' be the straight lines $t = 0$ and $t = 1$, respectively, defined on the strip $(t, x) \in [0, 1] \times \mathbf{R}$, where part of the flow of (2) lies. We consider the return map $h : L \to L'$ (when it is defined) as follows. If $y \in L$ then $h(y) = x(1, y)$, where $x(t, y)$ denotes the solution of (2) such that $x(0, y) = y$. Note that a periodic solution $x(t, y)$ satisfies $h(y) = y$. The *multiplicity of a limit cycle* $x(t, y)$ is the multiplicity of y as a zero of the function $h(y) - y$. Multiplicity of limit cycles for (2) is studied in [1], [16].

The main results that we prove are stated in the following theorems.

THEOREM A. *Suppose that* $A(t) \not\equiv 0$ *and does not change sign. Then the following hold.*

(a) *The sum of multiplicities of all limit cycles of* (2) *is at most* 3.

(b) *Table* 1 *shows a more precise distribution of the limit cycles* (2) *when* $A(t) \geqq 0$ (*the case* $A(t) \leqq 0$ *has associated the table obtained reversing the inequalities for c and d*).

Theorem A(a), as we said before, is already known. The new contribution consists of the additional information given in Table 1. The proof of the results stated in this table will use ideas similar to those of [7].

THEOREM B. *Assume* $B(t) \not\equiv 0$ *and does not change sign. Then the following hold*:

(a) *The sum of multiplicities of all limit cycles of* (2) *is at most* 3.

(b) *Table* 2 *shows a more precise distribution of the limit cycles of* (2) *when* $B(t) \geqq 0$ (*the case* $B(t) \leqq 0$ *has associated the table obtained reversing the inequalities for c*).

TABLE 1

Maximum number of limit cycles of equation (2) *when* $A(t) \geqq 0$. *Here* $c = \int_0^1 C(t)dt$, $d = \int_0^1 B(t)\, e^{\int_0^t C(s)\, ds}\, dt$.

	$c<0$	$c=0$			$c>0$		
		$d<0$	$d=0$	$d>0$	$d<0$	$d=0$	$d>0$
Maximum number of limit cycles in the half-strip $x>0$ taking into account their multiplicity	1	1	0	0	2	0	0
Multiplicity of the limit cycle $x=0$	1	2	3	2	1	1	1
Maximum number of limit cycles in the half-strip $x<0$ taking into account their multiplicity	1	0	0	1	0	0	2

Theorem B improves Proposition 2.3 of [17]. Its proof also uses the ideas utilized in the proof of Theorem A plus some geometrical results associated with the change of coordinates $x \to -x$.

The results of these theorems are the best ones in the following sense: The maximum number of limit cycles stated in the two tables are realizable for Abel equations when

either A or B does not change sign. It is enough to consider $A(t)$, $B(t)$, and $C(t)$ constant functions.

Table 2 could be improved by introducing a new parameter (similar to the parameter d in Table 1) that would give us a maximum number of limit cycles such that their sum in the whole strip was always at most 3. Unfortunately, we have not found this parameter.

TABLE 2

Maximum number of limit cycles of equation (2) *when* $B(t) \geqq 0$. *Here* $c = \int_0^1 C(t)\,dt$.

	$c<0$	$c=0$	$c>0$
Maximum number of limit cycles in the half-strip $x>0$ taking into account their multiplicity	2	1	1
Multiplicity of the limit cycle $x=0$	1	2	1
Maximum number of limit cycles in the half-strip $x<0$ taking into account their multiplicity	1	1	2
The sum of the multiplicities is at most	3	3	3

Similar results to those of Theorems A and B are not possible when we consider that $C(t)$ does not change sign. In fact, the example of an Abel equation with an arbitrary number of limit cycles, which we mentioned before, can be constructed with $C(t)$ a constant function, as was shown in [12]. That example is of the form

$$\frac{dx}{dt} = \varepsilon f(t)x^3 + a(t)x^2 + \delta x,$$

where $|\delta|$ is small, $a(t)$ is a polynomial of degree 1, and $f(t)$ is a polynomial of degree $2n$, and it can have at least $n+3$ limit cycles for suitable a and f.

In fact if we find a bound on the number of limit cycles of (2) with $C(t)$ a constant function in terms of A and B, we could give a bound on the number of limit cycles that a quadratic system has. Theorems A and B can be used in any way to study the limit cycles of planar polynomial vector fields with homogeneous nonlinearities. We consider two-dimensional autonomous systems of differential equations

$$\dot{x} = \lambda x - y + P_n(x, y), \qquad \dot{y} = x + \lambda y + Q_n(x, y), \tag{3}$$

where P_n and Q_n are real homogeneous polynomials of degree $n \geqq 2$. These systems for arbitrary $n \geqq 2$ have been studied in [2]-[5], [17]. When $n=2$ we have a subclass of quadratic systems which has been studied in [7], [8], [12]. System (3) with $P_n(x, y) = (ax+by)R_{n-1}(x, y)$ and $Q_n(x, y) = (cx+dy)R_{n-1}(x, y)$, where R_{n-1} is a homogeneous polynomial of degree $n-1$, has been studied in [6], [9]-[11].

System (3) in polar coordinates can be written in the form

$$\dot{r} = \lambda r + r^n f(\theta), \qquad \dot{\theta} = 1 + r^{n-1} g(\theta), \tag{4}$$

with

$$f(\theta) = \cos\theta P_n(\cos\theta, \sin\theta) + \sin\theta Q_n(\cos\theta, \sin\theta),$$

$$g(\theta) = \cos\theta Q_n(\cos\theta, \sin\theta) - \sin\theta P_n(\cos\theta, \sin\theta).$$

It is known that the periodic orbits surrounding the origin of system (4) do not intersect the curve $\dot{\theta} = 0$ (see the Appendix of [4]). Therefore, these periodic orbits

can be studied by making the transformation introduced by Cherkas [5], $T(r, \theta) = (\rho, \theta)$, where

$$\rho = r^{n-1}/(1 + r^{n-1} g(\theta)). \tag{5}$$

In the new coordinates (ρ, θ), system (4) becomes the following Abel equation

$$\frac{d\rho}{d\theta} = A(\theta)\rho^3 + B(\theta)\rho^2 + (n-1)\lambda\rho, \tag{6}$$

where $A = (n-1)g(\lambda g - f)$, and $B = (n-1)(f - 2\lambda g) - g'$.

In short, by studying all the periodic solutions $\rho(\theta)$ of (6) we study all the periodic solutions of system (3) surrounding the origin. Then by using Theorems A and B we can prove the following result.

THEOREM C. (a) *Suppose that either A or B does not change sign, $B \not\equiv 0$, and $A \not\equiv 0$. Then system* (3) *has at most two limit cycles surrounding the origin.*

(b) *If either $A \equiv 0$ or $B \equiv 0$ system* (3) *has at most one limit cycle surrounding the origin.*

Examples of system (3) with the maximum number of limit cycles given in the above theorem are given in [2], [3], [9]. For a more detailed study of the number of limit cycles of system (3), see Propositions 7–9 of § 4.

Note that if $B \not\equiv 0$ in (6), then it changes sign when n is even.

The rest of the paper is organized in the following way. In § 2 we state some auxiliary results that we will need in the proofs of Theorems A and B, which are given in § 3. Lastly, the cases $A \equiv 0$, $B \equiv 0$, and the proof of Theorem C are found in § 4.

2. Preliminary results. We will need the following results.

PROPOSITION 1 (see [15]). *If $h(y)$ is the return map associated with the differential equation $dx/dt = S(x, t)$, $0 \leqq t \leqq 1$, then*

$$\text{(a)}\quad h'(y) = \exp \int_0^1 \frac{\partial S}{\partial x}(x(t, y), t)\, dt,$$

$$\text{(b)}\quad h''(y) = h'(y)\left[\int_0^1 \frac{\partial^2 S}{\partial x^2}(x(t, y), t) \exp\left\{\int_0^t \frac{\partial S}{\partial x}(x(s, y), s)\, ds\right\} dt\right],$$

$$\text{(c)}\quad h'''(y) = h'(y)\left[\frac{3}{2}\left(\frac{h''(y)}{h'(y)}\right)^2 + \int_0^1 \frac{\partial^3 S}{\partial x^3}(x(t, y), t) \exp\left\{2\int_0^t \frac{\partial S}{\partial x}(x(s, y), s)\, ds\right\} dt\right],$$

where $x(t, y)$ denotes the solution of the differential equation such that $x(0, y) = y$.

LEMMA 2. *The first derivative of the return map associated with a periodic orbit $x(t)$ of* (2) *is*

$$\exp \int_0^1 C(t)\, dt \quad \text{if } x(t) \equiv 0,$$

or

$$\exp\left[-\int_0^1 \{B(t)x(t) + 2C(t)\}\, dt\right] = \exp \int_0^1 [A(t)x^2(t) - C(t)]\, dt \quad \text{if } x(t) \not\equiv 0.$$

Proof. For any periodic orbit $x(t)$ we know from Proposition 1 that the first derivative of the return map is

$$\exp\left(\int_0^1 (3Ax^2(t) + 2Bx(t) + C)\, dt\right),$$

so if $x(t) \equiv 0$, the lemma follows. If $x(t) \not\equiv 0$ then from (2) we know that

$$\frac{x'(t)}{x(t)} = Ax^2(t) + Bx(t) + C,$$

and integrating between zero and 1 we obtain

$$0 = \int_0^1 (Ax^2(t) + Bx(t) + C)\, dt. \tag{7}$$

Hence, by multiplying (7) by -3 or -2 and adding it to $\int_0^1 (3Ax^2(t) + 2Bx(t) + C)\, dt$, the lemma follows. □

LEMMA 3. *It is not restrictive in the study of the number of limit cycles of* (2) *to consider* $-B$ *instead of* B, *or* $-A$ *and* $-C$ *instead of* A *and* C, *respectively.*

Proof. By using one of the following three changes of variables, $(x, t) \to (-x, t)$, $(x, t) \to (x, 1-t)$, or $(x, t) \to (-x, 1-t)$, the lemma follows. □

LEMMA 4. *Solutions of* (2) *in the region* $x > 0$ (*respectively,* $x < 0$) *can be studied in the region* $y \equiv x^{-2} > 0$ *as solutions of the differential equation* (8) (*respectively,* (9)).

$$\frac{dy}{dt} = -2A(t) - 2B(t)y^{1/2} - 2C(t)y, \tag{8}$$

$$\frac{dy}{dt} = -2A(t) + 2B(t)y^{1/2} - 2C(t)y. \tag{9}$$

The proof follows easily.

LEMMA 5. *Assume that* $B(t) \geqq 0$ *and does not vanish identically. If* $x(t)$ *is a periodic orbit of* (2), *then the flow of* (2) *in the strip* $[0, 1] \times \mathbf{R}$ *moves upward across the curve* $(t, -x(t))$.

Proof. If $x(t)$ is a periodic orbit of (2), then $x'(t) = Ax^3(t) + Bx^2(t) + Cx(t)$. Hence the tangent of the curve $(t, -x(t))$ has the direction $(1, -Ax^3(t) - Bx^2(t) - Cx(t))$. Since we know that the vector field given by (2) at the point $(t, -x(t))$ is $(1, -Ax^3(t) + Bx^2(t) - Cx(t))$, the lemma follows. □

3. Proof of Theorems A and B.

Proof of Theorem A. By Lemma 3 we can assume that $A(t) \geqq 0$. Since for (2) $(\partial^3 S/\partial x^3)(t, x) = 6A(t) \geqq 0$, from Proposition 1 we know that $h'''(x) \geqq 0$ for all x for which h is defined. So, by Rolle's theorem, the maximum number of limit cycles of (2) taking into account their multiplicities is three.

To show that Table 1 is right we use more information about h. From Proposition 1 and Lemmas 2 and 3 we have $h(0) = 0$, $h'(0) = \exp c$, $h''(0) = 2dh'(0)$, where $c = \int_0^1 C(t)\, dt$ and

$$d = \int_0^1 B(t) \exp\left\{\int_0^t C(s)\, ds\right\} dt.$$

Furthermore

$$h'(x(0)) = \exp \int_0^1 (A(t)x^2(t) - C(t))\, dt \quad \text{when } x(0) \neq 0. \tag{10}$$

Assume now that $c < 0$. Then from (10) for any fixed point $x \neq 0$ of h, $h'(x) > 1$, and Table 1 follows.

Consider the case $c \geqq 0$. In this case we can assume that $d \geqq 0$, since the case $d < 0$ follows from this one and Lemma 3. We define $H(x) := h(x) - x$. For H we know that $H(0) = 0$, $H'(0) = e^c - 1 \geqq 0$, and $H''(0) = 2de^c \geqq 0$. Now we are going to prove that

there are no limit cycles in the half-strip $x>0$. Assume that $x=x_0$ gives the initial condition for the closest positive periodic solution to $x\equiv 0$. Then $H'(x_0)\leqq 0$ because $x\equiv 0$ is unstable. Hence from Rolle's theorem there exists y, $0<y<x_0$ such that $H'(y)=0$ and $H''(y)\leqq 0$. Note that conditions $H''(0)\geqq 0$ and $H''(y)\leqq 0$ with $0<y$ are in contradiction with the fact $H'''(x)=h'''(x)>0$. The rest of Table 1 follows from part (a) except when $c>0$ and $d=0$, in the half-strip $x<0$. Lemma 3 reduces this last case to the same case but in the half-strip $x>0$. □

Proof of Theorem B. From Lemma 4 we have that (2) is equivalent to either (8) or (9). By Lemma 3 we can take B of suitable sign, so that in the y coordinates the return map satisfies

$$h''(y)=\pm\frac{h'(y)}{2}\int_0^1 B(t)y^{-3/2}(t)\exp\left\{\int_0^t(\mp B(s)y^{-1/2}(s)-2C(s))\right\}dt>0.$$

Hence, by Rolle's theorem, we have proved that the sum of the multiplicities of the limit cycles of (2) in any half-strip, $x>0$, or $x<0$, is at most 2.

To show the final result we have to consider more information about the stability and relative position of the possible limit cycles.

Again by Lemma 3 it is not restrictive to consider $c\geqq 0$ and $B\geqq 0$. From Lemma 2 we have that for any initial condition x_0 of a periodic orbit of (2) in the half-strip $x>0$, $h'(x_0)=\exp(-\int_0^1(Bx_0(t)\,dt+2C(t))\,dt)<1$. Hence there is at most one limit cycle in this region. If $c=0$ the study in the half-strip $x<0$ follows in the same way.

So in order to finish the proof of this theorem it only remains to show that the maximum number of limit cycles in the whole strip is three, taking into account their multiplicities.

We consider the case $c>0$; the case $c=0$ follows in a similar way. Assume that there is a limit cycle with initial condition $x_0>0$ and two limit cycles (or a double one) with initial conditions $0>x_1\geqq x_2$. Assume that $x_1>x_2$. The case $x_1=x_2$ follows by using the same kind of arguments. From the results proved until now we know that the three limit cycles are hyperbolic and we know also their stabilities. So since the origin is a repellor limit cycle, we have by Lemma 5 that $x_0>|x_2|>|x_1|$, because x_0 has to be different from $|x_1|$ and $|x_2|$ and if $x_0<|x_2|$ then another positive limit cycle would exist between x_0 and $|x_2|$. But, again by Lemma 5, between $-x_0$ and x_2 system (2) would have another limit cycle and this is not possible. So, either the limit cycle with initial condition x_0 or the limit cycle with initial condition x_2 does not exist. □

4. Cases $A\equiv 0$, $B\equiv 0$, and proof of Theorem C. When either $A\equiv 0$ or $B\equiv 0$, (2) is of Bernoulli type, and it is well known how to integrate it. Hence in these cases we can know exactly the trajectories of all periodic solutions. Their initial conditions are given in the following lemma.

LEMMA 6. *Set*

$$c=\int_0^1 C(t)\,dt,\quad d=\int_0^1 B(t)\exp\left\{\int_0^t C(s)\,ds\right\}dt,\quad d'=2\int_0^1 A(t)\exp\left\{2\int_0^t C(s)\,ds\right\}dt.$$

Then the following hold.

(a) *If $A\equiv 0$ and $c=d=0$ all trajectories of* (2) *in a neighbourhood of $x\equiv 0$ are periodic.*

(b) *If $A\equiv 0$ and $|c|+|d|\neq 0$,* (2) *has at most two periodic solutions. Furthermore, these solutions are the solutions with initial conditions*

$$x(0)=0,\qquad x(0)=\frac{1-e^c}{d},$$

defined for all t between zero and 1.

(c) *If* $B \equiv 0$ *and* $c = d' = 0$, *all trajectories of* (2) *in a neighbourhood of* $x \equiv 0$ *are periodic.*

(d) *If* $B \equiv 0$ *and* $|c| + |d'| \neq 0$, *equation* (2) *has at most three closed solutions. Furthermore, these solutions are the solutions with initial condition,*

$$x(0) = 0, \qquad x(0) = \pm \sqrt{\frac{1 - e^{2c}}{d'}},$$

defined for all t *between zero and* 1.

The proof follows by direct computations.

Before applying Theorems A and B and the above result on the Abel equation (6) associated with system (3), we state some elementary results that can be found in [3] and [4]. Note that for the Abel equation (6) associated with (3) we are interested in 2π-periodic solutions. It is not difficult to translate all our results to this case. It is enough to consider instead of θ the new parameter $t := \theta/2\pi$.

(R1) In the region $\dot{\theta} > 0$ the flow of system (3) is diffeomorphic (preserving the orientation) to the flow of the Abel equation (6) contained in the half-cylinder R_1 defined by $0 \leqq \rho < 1/g(\theta)$ where this last inequality only works when $g(\theta) > 0$; see Fig. 1.

(R2) In the region $\dot{\theta} < 0$ the flow of system (3) is diffeomorphic (reversing the orientation) to the flow of (6) contained in the region $R_2 = \{\rho < 0\} \cap \{\rho < 1/g(\theta)$ when $g(\theta) < 0\}$; see Fig. 1.

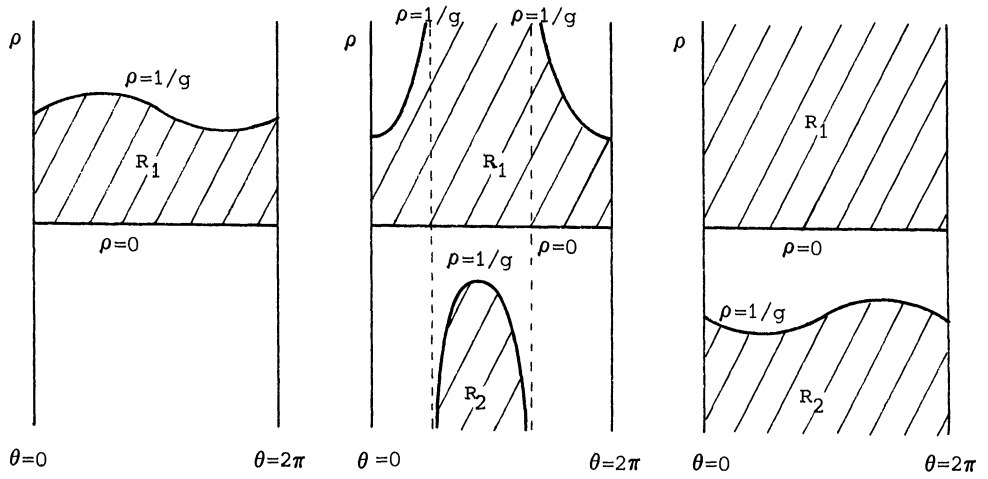


FIG. 1. *Some examples of regions* R_1 *and* R_2 *on the cylinder* (ρ, θ).

(R3) A periodic orbit of system (3) surrounding the origin is a periodic orbit of the Abel equation (6) contained in R_1 or R_2, and vice versa. Moreover, a periodic orbit can be contained in R_2 only if g is negative.

(R4) For the values of θ such that g does not vanish, the curve $\rho = 1/g$ is formed by solutions of the Abel equation (6).

(R5) If g does not vanish then the curve $\rho = 1/g$ is a periodic solution of the Abel equation (6).

(R6) The curve $\rho = 1/g$ for the Abel equation (6) corresponds to the equator of the Poincaré sphere of system (3) without the critical points.

(R7) Transformation T given in (5) sends the subsets $\dot{\theta} = 0$ to $\rho = \infty$, $r = 0$ to $\rho = 0$, and $r = \infty$ to $\rho = 1/g$.

(R8) If n is even and $\rho(\theta)$ is a solution of (6) then $-\rho(\theta+\pi)$ is also a solution of (6).

PROPOSITION 7. *Set* $d=\int_0^{2\pi}((n-1)(f-2\lambda g)-g')\,e^{(n-1)\lambda\theta}\,d\theta$, *and* $d'=2\int_0^{2\pi}(n-1)g(\lambda g-f)\,e^{2(n-1)\lambda\theta}\,d\theta$; *then the following hold.*

(a) *If* $\lambda g-f\equiv 0$ (*so* $A\equiv 0$) *and*

(1) $\lambda=0$, *then the origin of system* (3) *is a center.*

(2) $\lambda\neq 0$, *then system* (3) *has no limit cycles surrounding the origin.*

(b) *If* $g\equiv 0$ (*so* $A\equiv 0$) *and*

(1) $\lambda=d=0$, *then the origin of system* (3) *is a center.*

(2) *either* $\lambda=0$ *and* $d\neq 0$ *or* $\lambda\neq 0$, *then system* (3) *has at most one limit cycle surrounding the origin. Furthermore, if this limit cycle exists its initial condition is* $\rho(0)=(1-e^{\lambda(n-1)2\pi})/d$.

(c) *If* $(n-1)(f-2\lambda g)-g'\equiv 0$ *and*

(1) $\lambda=0$, *then the origin of system* (3) *is a center.*

(2) $\lambda\neq 0$, *then system* (3) *has at most one limit cycle surrounding the origin. Furthermore if this limit cycle exists its initial condition is* $\rho(0)=-\text{sign}\,(g(0))\sqrt{(1-e^{\lambda(n-1)4\pi})/d'}$ *and coincides with the function* $\rho(\theta)=-1/g(\theta)$.

Proof. Note that for (6) $c=\lambda(n-1)2\pi$. Hence in order to finish the proof of this proposition it suffices to show that the possible periodic solutions of (6) that Lemma 6 gives do not produce periodic orbits of system (3) except in the cases in which the origin is a center and in cases (b2) and (c2). Consider Case (a). In this case $f\equiv\lambda g$. So

$$d=-\int_0^{2\pi}((n-1)\lambda g+g')\,e^{\lambda(n-1)\theta}\,d\theta$$

$$=-g(\theta)\,e^{\lambda(n-1)\theta}\Big|_0^{2\pi}=g(0)(1-e^{\lambda(n-1)2\pi}).$$

Hence, by Lemma 6, the initial condition (different from zero) that gives us a possible limit cycle when $\lambda\neq 0$ is

$$x(0)=\frac{1-e^{\lambda(n-1)2\pi}}{g(0)(1-e^{\lambda(n-1)2\pi})}=\frac{1}{g(0)}.$$

Consequently, from (R6), case (a) follows. Case (b) follows, from Lemma 6, in a way similar to case (a). In case (c), again from Lemma 6, and with calculations similar to those in case (a), we have that the initial conditions that could give periodic orbits of (3) are $x(0)=\pm 1/g(0)$. So from (R6), the proposition is proved. □

PROPOSITION 8. (a) *If* $A(\theta)\not\equiv 0$, $A(\theta)$ *does not change sign and* n *is even, system* (3) *has at most one limit cycle surrounding the origin. Furthermore, it can exist only if* $c\cdot\text{sign}\,(A(\theta))<0$.

(b) *Assume* $A\not\equiv 0$, $A(\theta)\geqq 0$, *and that* n *is odd; then the following hold.*

(1) *If* $g(\theta)\equiv 0$ *then the number of limit cycles of system* (3) *surrounding the origin is at most the number appearing in the first row of Table* 1, *according to the signs of* c *and* d. *Furthermore, the limit cycles turn in the sense* $\dot\theta>0$.

(2) *If* $g(\theta)>0$ *for all* $\theta\in[0,2\pi]$, *then the number of limit cycles of system* (3) *surrounding the origin is at most the number appearing in the first row of Table* 1 *minus* 1, *according to the signs of* c *and* d. *Furthermore, they turn in the sense* $\dot\theta>0$.

(3) *If* $g(\theta)<0$ *for all* $\theta\in[0,2\pi]$ *system* (3) *has at most one limit cycle surrounding the origin. It can exist only if* $c<0$ *and then it turns in the sense* $\dot\theta>0$ *or if* $c>0$ *and* $d<0$ *and then it turns in the sense* $\dot\theta<0$.

The value c *is equal to* $\lambda(n-1)2\pi$ *and* d *is given in Proposition* 7. *If* $A(\theta)\leqq 0$ *we have similar results by reversing the inequalities for* c *and* d.

Proof. (a) The proof follows from Table 1 and results (R1), (R2), (R3), and (R8).

(b) The proof follows from Table 1 and results from (R1) to (R7). See also Fig. 1. Note that the case $A(\theta) \leqq 0$ can be obtained from case $A(\theta) \geqq 0$ by using Lemma 3. □

Most results of the two above propositions are already proved in [3].

PROPOSITION 9. *Assume that* $B(\theta) \not\equiv 0$ *and* $B(\theta) \geqq 0$ (*hence n is odd*). *Then the following hold.*

(a) *If* $g(\theta) \equiv 0$ *then the number of limit cycles of system* (3) *surrounding the origin is at most the number appearing in the first row of Table* 2, *according to the sign of c. Furthermore, the limit cycles turn in the sense* $\dot{\theta} > 0$.

(b) *If* $g(\theta) > 0$ *for all* $\theta \in [0, 2\pi]$, *then the number of limit cycles of system* (3) *surrounding the origin is at most the number appearing in the first row of Table* 2 *minus* 1, *according to the sign of c. Furthermore, they turn in the sense* $\dot{\theta} > 0$.

(c) *If* $g(\theta) < 0$ *for all* $\theta \in [0, 2\pi]$ *system* (3) *has at most one limit cycle surrounding the origin. It can exist only if* $c < 0$ *and then it turns in the sense* $\dot{\theta} > 0$ *or if* $c > 0$ *and* $d < 0$ *and then it turns in the sense* $\dot{\theta} < 0$.

The value c is equal to $\lambda(n-1)2\pi$. *If* $B(\theta) \leqq 0$ *we have similar results by reversing the inequalities for c.*

The proof of this proposition follows in a way similar to the proof of Proposition 8.

From Propositions 7-9 we obtain Theorem C.

Remark. Theorems A and B can also be applied to more general differential equations (not necessarily polynomial). It is enough that we can find a system of differential equations such that there exists a change of variables (usually polar coordinates) that transforms it into (2). So, for instance, we can apply Theorems A and B to a subclass of planar vector fields $X(v) = Cv + h(v)Dv$, studied in [9], where C and D are 2×2 matrices, h a smooth homogeneous function, when the functions $A(\theta)$ or $B(\theta)$ associated with this differential equation do not change sign.

REFERENCES

[1] M. A. M. ALWASH AND N. G. LLOYD, *Non autonomous equations related to polynomial two-dimensional systems*, Proc. Roy. Soc. Edinburgh Sect. A, 105 (1987), pp. 129-152.

[2] M. CARBONELL AND J. LLIBRE, *Limit cycles of a class of polynomial systems*, Proc. Roy. Soc. Edinburgh Sect. A, 109 (1988), pp. 187-199.

[3] ———, *Limit cycles of polynomial systems with homogeneous nonlinearities*, J. Math. Anal. Appl., 142 (1988), pp. 573-590.

[4] ———, *Hopf bifurcation, averaging methods and Liapunov quantities for polynomial systems with homogeneous nonlinearities*, Proc. European Conference on Iteration Theory-ECIT 87, World Scientific, Singapore, 1989, pp. 145-160.

[5] L. A. CHERKAS, *Number of limit cycles of an autonomous second-order system*, Differential Equations, 5 (1976), pp. 666-668.

[6] C. CHICONE, *Limit cycles of a class of polynomial vector fields in the plane*, J. Differential Equations, 63 (1986), pp. 68-87.

[7] B. COLL, A. GASULL, AND J. LLIBRE, *Some theorems on the existence, uniqueness, and nonexistence of limit cycles for quadratic systems*, J. Differential Equations, 67 (1987), pp. 372-399.

[8] W. A. COPPEL, *A simple class of quadratic systems*, J. Differential Equations, 64 (1986), pp. 275-282.

[9] A. GASULL, J. LLIBRE, AND J. SOTOMAYOR, *Limit cycles of vector fields of the form*: $X(v) = Av + f(v)Bv$, J. Differential Equations, 67 (1987), pp. 90-110.

[10] D. E. KODITSCHEK AND K. S. NARENDRA, *The stability of second order quadratic differential equations*, IEEE Trans. Automat. Control, AC-27(4) (1982), pp. 783-798.

[11] ———, *Limit cycles of planar quadratic differential equations*, J. Differential Equations, 54 (1984), pp. 181-195.

[12] A. LINSNETO, *On the number of solutions of the equation* $dx/dt = \sum_{j=0}^{n} a_j(t)x^j$, $0 \leqq t \leqq 1$ *for which* $x(0) = x(1)$, Invent. Math., 59 (1980), pp. 67-76.

[13] N. G. LLOYD, *The number of periodic solutions of the equation* $\dot{z} = z^N + p_1(t)z^{N-1} + \cdots + p_N(t)$, Proc. London Math. Soc., 27 (1973), pp. 667-700.

[14] ———, *On a class of differential equations of Riccati type*, J. London Math. Soc., 10 (1975), pp. 1-10.

[15] ———, *A note on the number of limit cycles in certain two-dimensional systems*, J. London Math. Soc., 20 (1979), pp. 277-286.

[16] ———, *Small amplitude limit cycles of polynomial differential equations*, Lecture Notes in Math. 1032, Springer-Verlag, Berlin, New York, 1983, pp. 346-357.

[17] ———, *Limit cycles of certain polynomial differential systems*, in Nonlinear Functional Analysis and Its Applications, S. P. Singh, ed., NATO AS1 Series C 173, D. Reidel, Dordrecht, the Netherlands, 1986, pp. 317-326.

[18] V. A. PLISS, *Non Local Problems of the Theory of Oscillations*, Academic Press, New York, 1966.

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1245-1263, September 1990
© 1990 Society for Industrial and Applied Mathematics
009

ON A SINGULARLY PERTURBED EIGENVALUE PROBLEM IN THE THEORY OF ELASTIC RODS*

L. S. FRANK†

Abstract. A singularly perturbed eigenvalue problem appearing in the theory of elastic rods is considered. The least eigenvalue λ_0^ε of the corresponding operator turns out to be exponentially decreasing as the small parameter ε vanishes, λ_0^ε being strictly positive for each $\varepsilon > 0$. Usual techniques based either on the parametrix constructions or on rescaling and stretching of variables fail to produce asymptotic formulae for λ_0^ε and the associated eigenfunction $\psi_0^\varepsilon(x)$ in the case considered. The classical geometrical optics approach is used here to derive asymptotic formulae for λ_0^ε and $\psi_0^\varepsilon(x)$ as $\varepsilon \to +0$.

Key words. singular perturbations, ellipticity, coerciveness, phase function, asymptotic expansions, ordinary differential equations, eigenvalue problems

AMS(MOS) subject classifications. 34E15, 34E20

0. Introduction. Singularly perturbed eigenvalue problems for ordinary differential operators affected by the presence of a small positive parameter is one of the classical topics in the singular perturbation theory that goes back to the work by Lord Rayleigh [12], where the following problem is considered:

$$\begin{aligned} &\varepsilon^2 u^{(iv)} - u'' = \lambda u, \qquad x \in (0,1), \\ &u(0) = u'(0) = u(1) = u'(1) = 0, \end{aligned} \tag{0.1}$$

and the following asymptotic formula for the eigenvalue λ_n^ε of this problem is established:

$$\lambda_n^\varepsilon = \pi^2 n^2 + 4\pi^2 n^2 \varepsilon + O(\varepsilon^2), \qquad \varepsilon \to 0, \quad n = 1, 2, \cdots. \tag{0.2}$$

The operator

$$L_\varepsilon := \varepsilon^2 \left(\frac{d}{dx}\right)^4 - \left(\frac{d}{dx}\right)^2, \qquad L_\varepsilon : D_{L_\varepsilon} \to L_2(0,1)$$

with the domain

$$D_{L_\varepsilon} := \{u \in H_4(0,1),\ u(0) = u'(0) = u(1) = u'(1) = 0\}$$

is self-adjoint for all $\varepsilon > 0$ and so it is for the reduced operator

$$L_0 = -\left(\frac{d}{dx}\right)^2, \qquad L_0 : D_{L_0} \to L_2(0,1)$$

with the domain

$$D_{L_0} = \{u \in H_2(0,1),\ u(0) = u(1) = 0\}.$$

Similar abstract self-adjoint problems are investigated in [9].

Not necessarily self-adjoint ordinary differential operators of the form

$$L_\varepsilon = \varepsilon^{2(n-m)} Q + P$$

with ord $Q = 2n >$ ord $P = 2m$ are considered in [11], where an assumption of strong ellipticity of L_ε (see [14]) is made and boundary conditions are considered that are a specific case of more general coercive boundary conditions for operators with a small parameter (see [1], [3]).

* Received by the editors June 16, 1989; accepted for publication November 9, 1989.
† Mathematics Department, Catholic University, Toernooiveld 6525, ED Nijmegen, the Netherlands.

Very essential in [11] is the assumption that the boundary operators associated with the reduced problem for $L_0 = P$ have their orders less than ord $P = 2m$.

The method used in [11] is closely related to that in [16] and goes back to the classical geometrical optics asymptotic method applied in the specific one-dimensional situation in [16].

The method introduced in [15] has the advantage of being applicable also in the case of elliptic partial differential operators. However, its realization in specific situations requires a considerable amount of technical work for deriving asymptotic expansions and for proving their convergence as the parameter vanishes.

The reduction method for coercive singular perturbations sketched in [2] and developed in [6] and [7] (see also [5], [8], [17]) allows us to derive, in a simple way, asymptotic formulae for the eigenvalues and eigenfunctions of coercive singular perturbations (see [4]) in the case where the perturbation shifts the spectrum of the reduced problem to distances whose order is some positive power of the small parameter. Since only the principal symbols of the coercive singular perturbations are used for producing a singularly perturbed operator that reduces a given coercive singular perturbation to a regular one, the reduction method based on such a construction cannot be applied in situations where the shift of the spectrum as a result of the perturbation is exponentially small when the parameter vanishes. Neither is the method in [15] applicable in this situation for the same reason as in the case of the reduction method mentioned above.

Yet problems of this type appear in a natural way in the theory of elastic rods. Such a problem is considered here and is analyzed directly by using the classical geometrical optics approach. The singularly perturbed eigenvalue problem in the interval $U = (0, 1)$ considered here is neither self-adjoint nor does it satisfy the conditions in [11], since for the corresponding reduced operator one of the boundary conditions on $\partial U = \{0, 1\}$ has the same order as that of the reduced operator in the interval U. Asymptotic formulae are derived and justified for the least eigenvalue and the associated eigenfunction of the coercive singular perturbation in the theory of elastic rods in the case where the rod is subjected to a large (rescaled dimensionless) longitudinal pulling-out force and has one of its endpoints clamped and the other free.

1. Statement of the problem. The following singularly perturbed boundary value problem describes an elastic rod at the equilibrium state in the presence of a large (rescaled dimensionless) pulling-out force when one of its endpoints is clamped and the other is free (see, for instance, [10]):

$$\varepsilon^2 D_x^2(q^2(x)D_x^2u(x)) + D_x^2u(x) = f(x), \qquad x \in U = (0, 1), \tag{1.1}$$

$$B_j(x', D_x)u(x') = \varphi_j(x'), \quad j = 1, 2, \quad x' \in \partial U = \{0, 1\}, \tag{1.2}$$

where $D_x = -id/dx$, $q(x) > 0$, for all $x \in U$, $q \in C^\infty(\bar{U})$, $\varepsilon \in (0, \varepsilon_0]$, $\varepsilon_0 \ll 1$ (ε is proportional to $T^{-1/2}$ with T the dimensionless parameter characterizing the pulling-out longitudinal force) and the boundary operators $B_j(x, D_x)$ are given as follows:

$$B_1(x, D_x) = (1-x) - xD_x^2, \qquad B_2(x, D_x) = B_1(x, D_x)\partial_n, \tag{1.3}$$

with $\partial_n = (-1)^{x'} d/dx$ the inward normal derivative.

We associate with (1.1)–(1.3) the following singularly perturbed column-operator:

$$\mathcal{A}^\varepsilon := (\pi_U r^\varepsilon D_x^2, \pi_{\partial U} B_1, \pi_{\partial U} B_2)^T, \tag{1.4}$$

where π_U and $\pi_{\partial U}$ are the restriction operators (traces of continuous functions) to $U = (0, 1)$ and $\partial U = \{0, 1\}$, respectively, the differential operator $r^\varepsilon(x, D_x)$ is defined as

$$r^\varepsilon(x, D_x) = \varepsilon^2 D_x^2 q^2(x) + 1, \tag{1.5}$$

and the upper T stands for the column vector that is the transpose of the corresponding row vector.

We associate with $\mathcal{A}^\varepsilon$ its reduced operator $\mathcal{A}^0$ defined as follows:

$$\mathcal{A}^0 := (\pi_U D_x^2,\ \pi_{\partial U} B_1)^T, \tag{1.6}$$

the corresponding reduced problem being stated in an obvious way:

$$\mathcal{A}^0 u^0 = (f, \varphi_1)^T. \tag{1.7}$$

Considering the boundary value problem

$$(\varepsilon^2 D_x^4 + D_x^2 - \lambda) u^\varepsilon(x) = 0, \qquad x \in U,$$

$$B_j(x', D_x) u(x') = \varphi_j(x'), \qquad j = 1, 2, \quad x' \in \partial U,$$

with $B_j(x, D_x)$ defined by (1.3) and with given $\lambda \in (0, \pi^2)$, it is readily seen that

$$\lim_{\varepsilon \to +0} u^\varepsilon(x) = u^0(x),$$

where $u^0(x)$ is the solution of the problem

$$(D_x^2 - \lambda) u^0(x) = 0, \qquad x \in U,$$

$$B_1(x', D_x) u^0(x') = \varphi_1(x'), \qquad x' \in \partial U.$$

Of course, the same is true if we consider inhomogeneous perturbed and reduced equations in U with a smooth second member $f(x)$ instead of zero.

This justifies the somewhat formal definition (1.6) of the reduced operator $\mathcal{A}^0$ associated with that of the perturbed operator defined by (1.4), (1.5), (1.3).

It is readily seen that the kernel of $\mathcal{A}^\varepsilon$ is trivial (while for the reduced operator it consists of all functions Cx with C any constant).

Indeed, introducing $v^\varepsilon = D_x^2 u^\varepsilon$ with $u^\varepsilon \in \ker \mathcal{A}^\varepsilon$, (i.e., $\mathcal{A}^\varepsilon u^\varepsilon = (0, 0, 0)^T$), for v^ε we get the following problem:

$$\begin{aligned} &r^\varepsilon v^\varepsilon = 0, \qquad x \in U, \\ &v^\varepsilon(1) = 0, \qquad D_x v^\varepsilon(1) = 0, \end{aligned} \tag{1.8}$$

so that $v^\varepsilon(x) \equiv 0$, for all $x \in \bar{U}$.

Furthermore, for u^ε we find

$$\begin{aligned} &D_x^2 u^\varepsilon = 0, \qquad x \in U, \\ &u^\varepsilon(0) = 0, \qquad D_x u^\varepsilon(0) = 0, \end{aligned} \tag{1.9}$$

so that $u^\varepsilon(x) \equiv 0$, for all $x \in \bar{U}$.

As a consequence of such a situation, it is impossible to factorize $\mathcal{A}^\varepsilon$ by $\mathcal{A}^0$, i.e., to find an operator R^ε (which would be a 3×2 matrix operator), such that $\mathcal{A}^\varepsilon = R^\varepsilon \mathcal{A}^0$, given that $\ker \mathcal{A}^\varepsilon = \{0\}$ and $\ker \mathcal{A}^0 = \{\text{Span } x\}$. Also, as (1.8) indicates, even if the inverse operator $(\mathcal{A}^\varepsilon)^{-1}$ exists, its norm grows exponentially, i.e., as $\exp(\gamma/\varepsilon)$ for $\varepsilon \to +0$ with some constant $\gamma > 0$. In other words, the eigenvalue $\lambda_0^0 = 0$ of the reduced operator (with x the associated eigenfunction) is shifted for its perturbation $\mathcal{A}^\varepsilon$ given by (4), to some value $\lambda_0^\varepsilon = O(\exp(-\gamma/\varepsilon))$ with some constant $\gamma > 0$, as $\varepsilon \to +0$.

We will find asymptotic formulae for λ_0^ε and the associated eigenfunction of $\mathcal{A}^\varepsilon$.

A direct method based on the classical geometrical optics approach will be applied for deriving these asymptotic formulae.

Thus, consider the eigenvalue problem

$$\pi_U(r^\varepsilon D_x^2-\lambda_0^\varepsilon)\psi_0^\varepsilon=0,\quad \pi_{\partial U}B_j(x,D_x)\psi_0^\varepsilon=0,\quad j=1,2, \tag{1.10}$$

where $B_j(x, D_x)$ are given by (1.3).

Remark 1.1. Singularly perturbed operator $\mathscr{A}^\varepsilon$ defined by (1.3)-(1.5) is a coercive singular perturbation (see [3]) and so it is also for the following perturbation $\mathscr{A}_\mu^\varepsilon$ of $\mathscr{A}^\varepsilon$:

$$\mathscr{A}_\mu^\varepsilon := (\pi_U(r^\varepsilon D_x^2+\mu^2),\ \pi_{\partial U}B_1,\ \pi_{\partial U}B_2)^T,\qquad \mu>0.$$

Since the reduced operator

$$\mathscr{A}_\mu^0 := (\pi_U(D_x^2+\mu^2),\ \pi_{\partial U}B_1)^T,\qquad \mu>0$$

is invertible, so it is also for $\mathscr{A}_\mu^\varepsilon$, for all $\varepsilon\in(0,\varepsilon_0]$, provided that $\varepsilon_0>0$ is sufficiently small, and, moreover, $(\mathscr{A}_\mu^\varepsilon)^{-1}$ is uniformly bounded with respect to $\varepsilon\in(0,\varepsilon_0]$ as a linear operator in corresponding Sobolev type spaces (see [5]-[7], [17]).

An unusual feature of $\mathscr{A}_\mu^\varepsilon$, $\mu\geqq 0$, is the boundary operators $B_k(x', D_x)$, $k=1,2$, whose orders are different at $x'=0$ and $x'=1$: $0=\operatorname{ord} B_1(0,D_x)<\operatorname{ord} B_1(1,D_x)=2$ and $1=\operatorname{ord} B_2(0,D_x)<\operatorname{ord} B_2(1,D_x)=3$.

2. Asymptotic solutions. Here two different types of asymptotic solutions to the differential equation in (1.10) with $\lambda\in\mathbf{C}$ instead of λ_0^ε will be constructed.

We denote by $q(x)$, $q:\mathbf{R}\to\mathbf{R}$ a smooth extension of $q\in C^\infty(\bar U)$ to $\mathbf{R}$ such that $q(x)\geqq q_0>0$, for all $x\in\mathbf{R}$ and $q(x)=q_\infty+q_1(x)$ with $q_1\in C_0^\infty(\mathbf{R})$ (see, for instance, [13], where the possibility of such an extension is shown).

We introduce the notation

$$L(\varepsilon,\lambda,x,\partial_x) := \varepsilon^2\partial_x^2q^2(x)\partial_x^2-\partial_x^2-\lambda, \tag{2.1}$$

where $\partial_x=d/dx$, $\varepsilon\in(0,\varepsilon_0]$, $\lambda\in\mathbf{C}$, $x\in\mathbf{R}$ and $q(x)$, $q\in C^\infty(\mathbf{R})$ is extended as indicated above.

We start with the following proposition.

PROPOSITION 2.1. *The equation*

$$L(\varepsilon,\lambda,x,\partial_x)u_\lambda^\varepsilon(x)=0,\qquad x\in\mathbf{R}, \tag{2.2}$$

has the following formal asymptotic solutions:

$$u_j(\varepsilon,\lambda,x)\sim\sum_{k\geqq 0}\varepsilon^{2k}u_{jk}(\lambda,x),\qquad j=1,2, \tag{2.3}$$

$$u_j(\varepsilon,\lambda,x)\sim\exp(-Q_j(x)/\varepsilon)\sum_{k\geqq 0}\varepsilon^k u_{jk}(\lambda,x),\qquad j=3,4, \tag{2.4}$$

where

$$u_{10}(\lambda,x)=\lambda^{-1/2}\sin(\lambda^{1/2}x),\qquad u_{20}(\lambda,x)=\cos(\lambda^{1/2}x), \tag{2.5}$$

$$Q_3(x)=\int_0^x(q(y))^{-1}\,dy,\quad Q_4(x)=\int_x^1(q(y))^{-1}\,dy,\quad u_{j0}(\lambda,x)=(q(x))^{1/2},\quad j=3,4, \tag{2.6}$$

and where $u_{jk}(\lambda,x)$, $k>0$, $1\leqq j\leqq 4$ *are defined recursively as follows*:

(i) *For* $j=1,2$

$$(\partial_x^2+\lambda)u_{jk}(\lambda,x)=\partial_x^2q^2\partial_x^2u_{j,k-1}(\lambda,x),\quad u_{jk}(\lambda,0)=\partial_xu_{jk}(\lambda,0)=0,\quad k=1,2,\cdots; \tag{2.7}$$

(ii) *For* $j=3,4$

$$
\begin{aligned}
&u_{jk}(\lambda, x)\equiv 0,\ k<0, \qquad u_{j0}(\lambda, x)=(q(x))^{1/2},\\
&L_q(u_{jk})(\lambda, x)=((L_q q(x)L_q+q(x)(\partial_x^2-\lambda))u_{j,k-1})(\lambda, x)\\
&\qquad -((L_q q^2(x)\partial_x^2+q^2(x)\partial_x^2 q(x)L_q)u_{j,k-2})(\lambda, x)\\
&\qquad +(q^2(x)\partial_x^2 q^2(x)\partial_x^2 u_{j,k-3})(\lambda, x), \qquad k=0,1,\cdots,\\
&u_{3,k}(\lambda, 0)=0,\ u_{4,k}(\lambda, 1)=0,\ k=1,2,\cdots,
\end{aligned}
\tag{2.8}
$$

with

$$
L_q(u)\coloneqq \partial_x u+q(x)\partial_x(u/q(x)). \tag{2.9}
$$

Proof. An elementary computation shows that for the formal asymptotic solutions $u_j(\varepsilon, \lambda, x)$, $j=1,2$, defined by (2.3), (2.7), holds:

$$
L(\varepsilon, \lambda, x, \partial_x)u_j(\varepsilon, \lambda, x)=O(\varepsilon^{\infty}), \qquad \varepsilon\to+0, \quad j=1,2.
$$

We show briefly that formally we have

$$
L(\varepsilon, \lambda, x, \partial_x)u_3(\varepsilon, \lambda, x)=\rho(\varepsilon, \lambda, x)\exp(-Q_3(x)/\varepsilon),
$$

where, uniformly with respect to $x\in\mathbf{R}$ and $|\lambda|\leqq r<\infty$,

$$
\rho(\varepsilon, \lambda, x)=O(\varepsilon^{\infty}), \qquad \varepsilon\to+0. \tag{2.10}
$$

Indeed, when we rewrite $L(\varepsilon, \lambda, x, \partial_x)$ in the form

$$
L(\varepsilon, \lambda, x, \partial_x)=\partial_x^2 r(x, \varepsilon\partial_x)-\lambda, \qquad r(x, \varepsilon\partial_x)=q^2(x)(\varepsilon\partial_x)^2-1, \tag{2.11}
$$

an elementary straightforward computation shows that

$$
r(x, \varepsilon\partial_x)u_3(\varepsilon, \lambda, x)\sim\varepsilon^2\exp(-Q_3(x)/\varepsilon)\sum_{k\geqq 0}\varepsilon^k(-q(x)L_q(u_{3,k+1})+q^2(x)\partial_x^2 u_{3,k}). \tag{2.12}
$$

Substitution of (2.12) into (2.2) yields

$$
\begin{aligned}
&L(\varepsilon, \lambda, x, \partial_x)u_3(\varepsilon, \lambda, x)\\
&\sim\exp(-Q_3(x)/\varepsilon)\sum_{k\geqq 0}\varepsilon^k((q(x))^{-2}v_{k+2}(\lambda, x)-(q(x))^{-1}L_q(v_{k+1})+\partial_x^2 v_k(\lambda, x)),
\end{aligned}
\tag{2.13}
$$

where we have denoted

$$
v_{k+2}(\lambda, x)\coloneqq q(x)(L_q(u_{3,k+1}))(\lambda, x)+q^2(x)\partial_x^2 u_{3,k}(\lambda, x), \qquad k=0,1,\cdots. \tag{2.14}
$$

Using (2.14), (2.8), it is readily seen that

$$
(q(x))^{-2}v_{k+2}(\lambda, x)-(q(x))^{-1}(L_q(v_{k+1}))(\lambda, x)+\partial_x^2 v_k(\lambda, x)\equiv 0, \qquad k\geqq 0, \tag{2.15}
$$

where, of course, $v_k(\lambda, x)\equiv 0$, $k=0$, 1, as a consequence of (2.14), (2.8), and (2.6), the latter having as a consequence the identity $(L_q(u_{j,0}))(x)\equiv 0$. The same argument applies to $u_4(\varepsilon, \lambda, x)$. □

We will use the classical construction due to Carleman to produce functions $v_j(\varepsilon, \lambda, x)$ that are C^{∞} in variables $(\varepsilon, x)\in(0, \varepsilon_0]\times\bar{U}$, analytic in $\lambda\in\mathbf{C}$, and have the following properties: $v_j(\varepsilon, \lambda, x)$, $1\leqq j\leqq 4$, admit an asymptotic expansion by the formal series representing the corresponding $u_j(\varepsilon, \lambda, x)$, $1\leqq j\leqq 4$.

LEMMA 2.2. *Let* $\varphi_k\in C^{\infty}(\bar{U})$, $k\geqq 0$ *and* $\mu\in[\mu_0, \infty)$. *Then there exists a function* $\varphi(\mu, x)$, $\varphi\in C^{\infty}([\mu_0, \infty)\times U)$ *such that the following asymptotic relations hold*:

$$
\partial_\mu^\alpha\partial_x^\beta\varphi(\mu, x)\sim\sum_{k\geqq 0}(\partial_\mu^\alpha\mu^{-k})\partial_x^\beta\varphi_k(x), \qquad \mu\to+\infty \quad \forall\alpha, \beta, \tag{2.16}
$$

uniformly with respect to $x \in \bar{U}$, i.e., for each integer $N > 0$ we have

$$\left| \partial_\mu^\alpha \partial_x^\beta \varphi(\mu, x) - \sum_{0 \leqq k < N} (\partial_\mu^\alpha \mu^{-k}) \partial_x^\beta \varphi_k(x) \right| \leqq C_{\alpha,\beta,N} \mu^{-(N+\alpha)}, \tag{2.17}$$

where the constant $C_{\alpha,\beta,N}$ may depend only on its subscripts.

Proof. Let $\chi \in C^\infty(\mathbf{R})$ be such that $\chi(t) \equiv 1$ for $|t| \geqq 1$ and $\chi(t) \equiv 0$ for $|t| \leqq \frac{1}{2}$. Define

$$\varphi(\mu, x) = \sum_{k \geqq 0} \chi(\delta_k \mu) \mu^{-k} \varphi_k(x), \tag{2.18}$$

where the sequence $\delta_k \downarrow 0$ for $k \to \infty$ will be chosen later.

Note that, for each $\mu \geqq \mu_0$ given, the series on the right-hand side of (2.18) contains only a finite number nonvanishing terms, i.e., it is convergent for each given $\mu \geqq \mu_0$. The numbers $\delta_k \downarrow 0$, $k \to \infty$, are chosen to satisfy

$$|\partial_\mu^\alpha \partial_x^\beta (\chi(\delta_k \mu) \mu^{-k} \varphi_k(x))| \leqq \mu^{1-\alpha-k}, \quad 0 \leqq \alpha + \beta \leqq k \quad \forall \mu \geqq \mu_0, \quad \forall x \in \bar{U}. \tag{2.19}$$

Thus, for all $N > 0$ and for $\alpha + \beta \leqq N$, we find

$$\left| \sum_{k \geqq N+2} \partial_\mu^\alpha \partial_x^\beta (\chi(\delta_k \mu) \mu^{-k} \varphi_k(x)) \right| \leqq \sum_{k \geqq N+1} \mu^{-1-\alpha-k} = O(\mu^{-1-\alpha-N}) \quad \text{as } \mu \to \infty. \tag{2.20}$$

Thus (2.20) yields

$$\begin{aligned} \partial_\mu^\alpha \partial_x^\beta \varphi(\mu, x) &= \partial_\mu^\alpha \partial_x^\beta \sum_{0 \leqq k \leqq N+1} \chi(\delta_k \mu) \mu^{-k} \varphi_k(x) + O(\mu^{-1-\alpha-N}) \\ &= \partial_\mu^\alpha \partial_x^\beta \sum_{0 \leqq k \leqq N+1} \mu^{-k} \varphi_k(x) + O(\mu^{-1-\alpha-N}) \quad \text{as } \mu \to \infty. \end{aligned}$$

□

Now define the functions

$$v_j(\varepsilon, \lambda, x) = \sum_{k \geqq 0} \chi(\delta_k \varepsilon^{-2}) \varepsilon^{2k} u_{jk}(\lambda, x), \qquad j = 1, 2, \tag{2.21}$$

$$v_j(\varepsilon, \lambda, x) = \exp(-Q_j(x)/\varepsilon) \sum_{k \geqq 0} \chi(\delta_k \varepsilon^{-1}) \varepsilon^k u_{jk}(\lambda, x), \qquad j = 3, 4, \tag{2.22}$$

where u_{jk}, $j = 1, 2$ and u_{jk}, $j = 3, 4$ are defined by (2.7) and (2.8), respectively.

THEOREM 2.3. *Equation* (2.2) *has a fundamental system of solutions $w_j(\varepsilon, \lambda, x)$, $1 \leqq j \leqq 4$ such that*:

(i) *w_j are C^∞ in variables $(\varepsilon, x) \in (0, \varepsilon_0] \times \bar{U}$ and analytic in variable $\lambda \in \mathbf{C}$;*

(ii) *For any $\alpha \geqq 0$, $\beta \geqq 0$, and $N > 0$ given the following asymptotic relations hold*:

$$(\varepsilon^3 \partial_\varepsilon)^\alpha \partial_x^\beta (w_j(\varepsilon, \lambda, x) - v_j(\varepsilon, \lambda, x)) = O(\varepsilon^{2\alpha+2N}), \qquad j = 1, 2, \tag{2.23}$$

$$(\varepsilon^2 \partial_\varepsilon)^\alpha \partial_x^\beta (w_j(\varepsilon, \lambda, x) - v_j(\varepsilon, \lambda, x)) = O(\varepsilon^{\alpha+N} \exp(-Q_j(x)/\varepsilon), \qquad j = 1, 2, \tag{2.24}$$

or, equivalently,

$$(\varepsilon^2 \partial_\varepsilon)^\alpha \partial_x^\beta (\exp(Q_j(x)/\varepsilon)(w_j(\varepsilon, \lambda, x) - v_j(\varepsilon, \lambda, x))) = O(\varepsilon^{\alpha+N}), \tag{2.25}$$

the asymptotic formulae (2.23), (2.25) *being valid uniformly with respect to $x \in \bar{U}$ and $\lambda \in \mathbf{C}$, $|\lambda| \leqq r < \infty$.*

Proof. We briefly sketch the proof emphasizing the main ideas and constructions and omitting technical details that can be easily verified.

Let us start with the solutions $w_j(\varepsilon, \lambda, x)$, $j=1,2$ of the first type. Using a smooth extension of q onto $\mathbf{R}$ (with the properties indicated above), we may consider the equation

$$L(\varepsilon, \lambda, x, \partial_x) w_j(\varepsilon, \lambda, x) = 0, \qquad x \in \mathbf{R}, \quad j=1,2. \tag{2.26}$$

As a consequence of Proposition 2.1 and Lemma 2.2, we have for the asymptotic solutions $v_j(\varepsilon, \lambda, x)$, $j=1,2$ defined by (2.3), (2.7), (2.21)

$$L(\varepsilon, \lambda, x, \partial_x) v_j(\varepsilon, \lambda, x) = \delta_j(\varepsilon, \lambda, x), \qquad j=1,2, \tag{2.27}$$

where $\delta_j(\varepsilon, \lambda, x)$ is $O(\varepsilon^\infty)$, as $\varepsilon \to 0$, with all its derivatives, i.e., in the topology of $C^\infty((0, \varepsilon_0] \times \mathbf{R})$; furthermore, as a consequence of the construction, $\lambda \to \delta_j(\varepsilon, \lambda, x)$ is an entire function of $\lambda \in \mathbf{C}$.

We will seek the solution w_j of (2.26) such that

$$w_j(\varepsilon, \lambda, 0) = v_j(\varepsilon, \lambda, 0), \quad \partial_x w_j(\varepsilon, \lambda, 0) = \partial_x v_j(\varepsilon, \lambda, 0), \quad j=1,2. \tag{2.28}$$

Introducing $Y_j(\varepsilon, \lambda, x) = v_j(\varepsilon, \lambda, x) - w_j(\varepsilon, \lambda, x)$, $j=1,2$, we find

$$\begin{aligned} &(r^\varepsilon(x, D_x) D_x^2 - \lambda) Y_j(\varepsilon, \lambda, x) = \delta_j(\varepsilon, \lambda, x), \qquad x \in \mathbf{R}, \\ &Y_j(\varepsilon, \lambda, 0) = 0, \quad \partial_x Y_j(\varepsilon, \lambda, 0) = 0, \quad j=1,2, \end{aligned} \tag{2.29}$$

where $r^\varepsilon(x, D_x) = \varepsilon^2 D_x^2 q^2(x) + 1$ is an elliptic singular perturbation of order $(0, 0, 2)$.

As a consequence of Theorem 2.2.1 in [6], $r^\varepsilon(x, D_x)$ has an inverse $s^\varepsilon(x, D_x)$ that is an elliptic singular perturbation of order $(0, 0, -2)$. Hence we may rewrite (2.29) equivalently in the following fashion:

$$Y_j(\varepsilon, \lambda, x) - \lambda \int_0^x (x-y) s^\varepsilon(y, D_y) Y_j(\varepsilon, \lambda, y)\, dy = \rho_j(\varepsilon, \lambda, x), \tag{2.30}$$

with

$$\rho_j(\varepsilon, \lambda, x) = \int_0^x (x-y) s^\varepsilon(y, D_y) \delta_j(\varepsilon, \lambda, y)\, dy, \qquad j=1,2$$

still being $O(\varepsilon^\infty)$ in the topology of $C^\infty_{\varepsilon,x}$ and entire functions of $\lambda \in \mathbf{C}$.

Volterra integral equations (2.30) have well-defined solutions $Y_j(\varepsilon, \lambda, x)$, $j=1,2$, that possess the same properties, as $\rho_j(\varepsilon, \lambda, x)$, i.e.,

$$Y_j(\varepsilon, \lambda, x) = O(\varepsilon^\infty), \qquad j=1,2, \tag{2.31}$$

in the topology of $C^\infty((0, \varepsilon_0] \times \mathbf{R})$ and are entire functions of $\lambda \in \mathbf{C}$.

Next, we consider the solutions $w_j(\varepsilon, \lambda, x)$, $j=3,4$ of the second type. Again, using a smooth extension of $q(x)$ onto $\mathbf{R}$ (as indicated above), we may consider the equation

$$L(\varepsilon, \lambda, x, \partial_x) w_j(\varepsilon, \lambda, x) = 0, \qquad x \in \mathbf{R}, \quad j=3,4. \tag{2.32}$$

Again, as a consequence of Proposition 2.1 and Lemma 2.2, we have for the asymptotic solutions $v_j(\varepsilon, \lambda, x)$ defined by (2.4), (2.5), (2.8), (2.22)

$$\exp(Q_j(x)/\varepsilon) L(\varepsilon, \lambda, x, \partial_x) v_j(\varepsilon, \lambda, x) = \delta_j(\varepsilon, \lambda, x), \qquad j=3,4, \tag{2.33}$$

where again $\delta_j(\varepsilon, \lambda, x) = O(\varepsilon^\infty)$ in the topology of $C^\infty_{\varepsilon,x}$ and are entire functions of $\lambda \in \mathbf{C}$.

We will seek solutions $w_j(\varepsilon, \lambda, x)$ of (2.32) that satisfy the conditions

$$w_j(\varepsilon, \lambda, 0) = v_j(\varepsilon, \lambda, 0), \qquad j=3,4,$$

so that for

$$Y_j(\varepsilon, \lambda, x) := \exp(Q_j(x)/\varepsilon)(v_j(\varepsilon, \lambda, x) - w_j(\varepsilon, \lambda, x)), \tag{2.34}$$

we get the following problem:

$$\begin{aligned} &\exp(Q_j(x)/\varepsilon)L(\varepsilon,\lambda,x,\partial_x)(\exp(-Q_j(x)/\varepsilon)Y_j(\varepsilon,\lambda,x))=\delta_j(\varepsilon,\lambda,x), \qquad x\in\mathbf{R},\\ &Y_j(\varepsilon,\lambda,0)=0, \qquad j=3,4. \end{aligned} \tag{2.35}$$

A straightforward computation shows that

$$\varepsilon\exp(Q_j(x)/\varepsilon)L(\varepsilon,\lambda,x,\partial_x)\exp(-Q_j(x)/\varepsilon)=M(x,\varepsilon D_x)\partial_x+R(\varepsilon,\lambda,x,\partial_x), \tag{2.36}$$

where

$$M(x,\eta):=-iq^2(x)\eta^3+4q(x)\eta^2+5i\eta-6/q(x), \tag{2.37}$$

$M(x,\varepsilon D_x)$ is an elliptic perturbation of order $(0,0,3)$ and $R(x,\varepsilon,\partial)$ is a differential singular perturbation whose order is at most $(0,0,2)$.

Hence, again as a consequence of Theorem 2.2.1 in [6], $M(x,\varepsilon D_x)$ has an inverse $N(x,\varepsilon,D_x)$, which is an elliptic singular perturbation of order $(0,0,-3)$.

Therefore we may rewrite (2.35) in the following equivalent fashion:

(2.38)

$$Y_j(\varepsilon,\lambda,x)+\int_0^x N(\varepsilon,\lambda,D_y)\circ R(\varepsilon,\lambda,y,\partial_y)Y_j(\varepsilon,\lambda,y)\,dy=\rho_j(\varepsilon,\lambda,x), \qquad j=3,4,$$

where $\rho_j(\varepsilon,\lambda,x)$ have the same properties as $\delta_j(\varepsilon,\lambda,x)$.

Thus the well-defined solutions Y_j of Volterra integral equations with

$$\operatorname{ord} N(\varepsilon,\lambda,D_y)\circ R(\varepsilon,\lambda,y,\partial_y)\leqq(0,0,-1)$$

have the same properties as $\delta_j(\varepsilon,\lambda,x)$, i.e.,

$$Y_j(\varepsilon,\lambda,x)=O(\varepsilon^\infty), \qquad j=3,4,$$

in the topology of $C^\infty((0,\varepsilon_0]\times\mathbf{R})$ and are entire functions of $\lambda\in\mathbf{C}$. □

Remark 2.4. The results in Theorem 2.3 are similar to those in [11] and [16]. However, their proof given here and based on the reduction theory developed in [6] and [7] is quite different and much simpler.

3. Asymptotic formulae. Here asymptotic formulae for the eigenvalues and the eigenfunctions of the operator $\mathscr{A}^\varepsilon$ defined by (1.3)-(1.5) are derived by means of Proposition 2.1 and Theorem 2.3.

Let $w_j(\varepsilon,\lambda,x)$, $1\leqq j\leqq 4$, be the fundamental system of solutions of (2.2), which has been constructed in the proof of Theorem 2.3. Introduce the matrix

$$\mathscr{D}(\varepsilon,\lambda):=\begin{Vmatrix} w_1(\varepsilon,\lambda,0), & w_2(\varepsilon,\lambda,0), & \varepsilon w_3(\varepsilon,\lambda,0), & \varepsilon^3 w_4(\varepsilon,\lambda,0)\\ w_1'(\varepsilon,\lambda,0), & w_2'(\varepsilon,\lambda,0), & \varepsilon w_3'(\varepsilon,\lambda,0), & \varepsilon^3 w_4'(\varepsilon,\lambda,0)\\ w_1''(\varepsilon,\lambda,1), & w_2''(\varepsilon,\lambda,1), & \varepsilon w_3''(\varepsilon,\lambda,1), & \varepsilon^3 w_4''(\varepsilon,\lambda,1)\\ w_1'''(\varepsilon,\lambda,1), & w_2'''(\varepsilon,\lambda,1), & \varepsilon w_3'''(\varepsilon,\lambda,1), & \varepsilon^3 w_4'''(\varepsilon,\lambda,1) \end{Vmatrix}, \tag{3.1}$$

where the upper dash stands for the derivative with respect to x: $'=d/dx$.

Obviously, the eigenvalues of $\mathscr{A}^\varepsilon$ defined by (1.3)-(1.5) are the zeros of

$$F(\varepsilon,\lambda):=\det\mathscr{D}(\varepsilon,\lambda)=0, \qquad \varepsilon\in(0,\varepsilon_0]. \tag{3.2}$$

$F(\varepsilon,\lambda)$ being an entire function of $\lambda\in\mathbf{C}$ for each $\varepsilon\in(0,\varepsilon_0]$, then (3.2) has isolated zeros λ_n^ε, $n=0,1,\cdots$ such that $|\lambda_n^\varepsilon|\to\infty$ for $n\to\infty$.

As a consequence of Proposition 2.1 and Theorem 2.3, we have

$$
\begin{aligned}
F(0,\lambda) &:= \lim_{\varepsilon\to+0} F(\varepsilon,\lambda) \\
&= \det \begin{Vmatrix} O, & 1, & O, & O \\ 1, & O, & -(q(0))^{-1/2}, & O \\ -\lambda^{1/2}\sin\lambda^{1/2}, & -\lambda\cos\lambda^{1/2}, & O, & O \\ -\lambda\cos\lambda^{1/2}, & \lambda^{3/2}\sin\lambda^{1/2}, & O, & (q(1))^{-5/2} \end{Vmatrix} \qquad (3.3) \\
&= (q(0))^{-1/2}(q(1))^{-5/2}\lambda^{1/2}\sin\lambda^{1/2}.
\end{aligned}
$$

The zeros

$$
\lambda_n^0 = \pi^2 n^2, \qquad n = 0, 1, \cdots \tag{3.4}
$$

of $F(0,\lambda)$ are nothing but the eigenvalues of the reduced operator $\mathscr{A}^0$ defined by (1.6), (1.3).

LEMMA 3.1. *The following asymptotic formula holds uniformly with respect to λ on each compact set in* $\mathbf{C}$:

$$
\begin{aligned}
F(\varepsilon,\lambda) = F(0,\lambda) - \varepsilon(q(0))^{-1/2}(q(1))^{-5/2}\Big(&\lambda\cos\lambda^{1/2}\sum_{x'\in\partial U}\pi_{\partial U}q(x') \\
&+\lambda^{1/2}\sin\lambda^{1/2}\sum_{x'\in\partial U}\pi_{\partial U}\partial_n q(x')\Big) + O(\varepsilon^2),
\end{aligned} \tag{3.5}
$$

where $\partial_n = (-1)^{x'} d/dx$ is the inward normal derivative at $x' \in \partial U$.

Proof. As a consequence of Proposition 2.1 and Theorem 2.3, we have the following asymptotic relations uniformly with respect to $x \in \bar{U}$ and λ belonging to each compact set in $\mathbf{C}$:

$$
\begin{aligned}
w_1(\varepsilon,\lambda,x) &= \lambda^{-1/2}\sin(\lambda^{1/2}x) + O(\varepsilon^2), \\
w_2(\varepsilon,\lambda,x) &= \cos(\lambda^{1/2}x) + O(\varepsilon^2), \\
w_3(\varepsilon,\lambda,x) &= \exp\left(-\varepsilon^{-1}\int_0^x (q(y))^{-1}\,dy\right)((q(x))^{1/2} + \varepsilon u_{31}(\lambda,x) + O(\varepsilon^2)), \qquad (3.6) \\
w_4(\varepsilon,\lambda,x) &= \exp\left(-\varepsilon^{-1}\int_x^1 (q(y))^{-1}\,dy\right)((q(x))^{1/2} + \varepsilon u_{41}(\lambda,x) + O(\varepsilon^2)),
\end{aligned}
$$

where $u_{j1}(\lambda,x)$, $j=3,4$, are defined by (2.8).

Inserting (3.6) into (3.1) and using (3.3), we get (3.5), while keeping all the terms with ε^k, $k=0,1$ and neglecting those that are $O(\varepsilon^2)$, as $\varepsilon\to+0$. □

THEOREM 3.2. *The eigenvalues λ_n^ε, $n=0,1,\cdots$ of $\mathscr{A}^\varepsilon$ defined by* (1.3)–(1.5) *are real and for each $n=0,1,\cdots$ the following asymptotic formulae are valid*:

$$
\lambda_n^\varepsilon = \pi^2 n^2\Big(1 - 2\varepsilon\sum_{x'\in\partial U}\pi_{\partial U}q(x')\Big) + O(\varepsilon^2). \tag{3.7}
$$

Proof. As a consequence of (3.3), we have for $F_\lambda = (\partial/\partial\lambda)F$:

$$
\begin{aligned}
&F_\lambda(0,\pi^2n^2) = (\tfrac{1}{2})(-1)^n(q(0))^{-1/2}(q(1))^{-5/2} \neq 0, \qquad n = 1,2,\cdots, \\
&F_\lambda(0,0) = (q(0))^{-1/2}(q(1))^{-5/2} \neq 0.
\end{aligned} \tag{3.8}
$$

Furthermore, as a consequence of (3.5), $F_\varepsilon(\varepsilon, \lambda) = (\partial/\partial\varepsilon)F(\varepsilon, \lambda)$ is continuous on $[0, \varepsilon_0] \times \mathbf{C}$ for each $\varepsilon_0 > 0$ and $F_{\varepsilon\varepsilon}(\varepsilon, \lambda) = (\partial/\partial\varepsilon)^2 F(\varepsilon, \lambda)$ is bounded on $(0, \varepsilon_0] \times K$ for each $\varepsilon_0 > 0$ and each compact $K \in \mathbf{C}$.

Since $\lambda_n^0 = \pi^2 n^2$, $n = 0, 1, \cdots$ are real and $\lambda_n^0 > 0$, $n = 1, 2, \cdots$, the implicit functions theorem implies that the zeros $\lambda_n^\varepsilon = \lambda_n(\varepsilon)$ of $F(\varepsilon, \lambda)$ defined by (3.2) are real and continuously differentiable with respect to $\varepsilon \in [0, \varepsilon_0]$ with any $\varepsilon_0 > 0$; moreover, for $\varepsilon_0 > 0$ sufficiently small we have: $\lambda_n^\varepsilon > 0$, $n = 1, 2, \cdots$, for all $\varepsilon \in [0, \varepsilon_0]$.

Furthermore, for μ_n,

$$\mu_n := \left(\frac{d}{d\varepsilon}\right)\lambda_n(\varepsilon)|_{\varepsilon=0}, \qquad n = 0, 1, \cdots \tag{3.9}$$

using (3.5) we find

$$\mu_n = -F_\varepsilon(0, \pi^2 n^2)(F_\lambda(0, \pi^2 n^2))^{-1} = -2\pi^2 n^2 \sum_{x' \in \partial U} \pi_{\partial U} q(x'), \qquad n = 0, 1, \cdots$$

and that proves (3.7).

Remark 3.3. In fact, using Proposition 2.1 and Theorem 2.3, we conclude that $F(\varepsilon, \lambda)$ defined by (3.1), (3.2) is infinitely differentiable with respect to $\varepsilon \geqq 0$, being an entire function of $\lambda \in \mathbf{C}$ with all its derivatives with respect to $\varepsilon \geqq 0$. Thus, the eigenvalues $\lambda_n^\varepsilon = \lambda_n(\varepsilon)$, $n = 0, 1, \cdots$ of $\mathscr{A}^\varepsilon$ defined by (1.3)-(1.5) are C^∞ in $\varepsilon \geqq 0$, and for each integer $N > 0$ and each $n = 1, 2, \cdots$ we have

$$\lambda_n(\varepsilon) = \sum_{0 \leqq k < N} \lambda_n^k \varepsilon^k + O(\varepsilon^N), \qquad \varepsilon \to +0,$$

where, of course,

$$\lambda_n^0 = \pi^2 n^2, \qquad \lambda_n^1 = -2\pi^2 n^2 \sum_{x' \in \partial U} \pi_{\partial U} q(x'),$$

and all λ_n^k, $k \geqq 2$ can be defined recursively using Proposition 2.1.

Furthermore, obviously, $\lambda_0^\varepsilon = \lambda_0(\varepsilon) = O(\varepsilon^N)$, for all $N > 0$, and an asymptotic formula for λ_0^ε (which is exponentially small as $\varepsilon \to +0$) will be exhibited later.

Now an asymptotic formula for $\lambda_0^\varepsilon = \lambda_0(\varepsilon)$ will be derived.

Using (2.7), for $F(\varepsilon, \lambda)$ defined by (3.1), (3.2) we find

$$F(\varepsilon, 0) = \varepsilon^4 (w_3'''(\varepsilon, 0, 1) w_4''(\varepsilon, 0, 1) - w_3''(\varepsilon, 0, 1) w_4'''(\varepsilon, 0, 1)), \tag{3.10}$$

and Proposition 2.1 and Theorem 2.3 yield, with any integer $N > 0$,

$$F(\varepsilon, 0) = -2(q(1))^{-4} \delta_\varepsilon \left(\sum_{0 \leqq k < N} \gamma_k \varepsilon^{k-1} + O(\varepsilon^N) \right), \qquad \varepsilon \to 0, \tag{3.11}$$

where $\gamma_0 = 1$ and γ_k, $k > 0$, are computed recursively using (2.8), and where we have denoted

$$\delta_\varepsilon := \exp\left(-\varepsilon^{-1} \int_0^1 (q(y))^{-1} \, dy \right). \tag{3.12}$$

The same argument yields for $F_\lambda(\varepsilon, 0)$

$$
\begin{aligned}
F_\lambda(\varepsilon,0) = \det \begin{Vmatrix} 0, & 1, & \varepsilon w_3(\varepsilon,0,0), & 0 \\ 0, & 0, & \varepsilon w_3'(\varepsilon,0,0), & 0 \\ -1, & 0, & 0, & \varepsilon^3 w_4''(\varepsilon,0,1) \\ -1, & 0, & 0, & \varepsilon^3 w_4'''(\varepsilon,0,1) \end{Vmatrix} \\
+ \det \begin{Vmatrix} 0, & 0, & \varepsilon w_3(\varepsilon,0,0), & 0 \\ 1, & 0, & \varepsilon w_3'(\varepsilon,0,0), & 0 \\ 0, & -1, & 0, & \varepsilon^3 w_4''(\varepsilon,0,1) \\ 0, & 0, & 0, & \varepsilon^3 w_4'''(\varepsilon,0,1) \end{Vmatrix} + O(\varepsilon^{-2}\delta_\varepsilon),
\end{aligned} \tag{3.13}
$$

since, as a consequence of (3.6), for $\varepsilon \to +0$ we have

$$
\partial_\lambda^k \partial_x^p w_3(\varepsilon,0,1) = O(\varepsilon^{-p}\delta_\varepsilon), \quad \partial_\lambda^k \partial_x^p w_4(\varepsilon,0,0) = O(\varepsilon^{-p}\delta_\varepsilon) \quad \forall p \geqq 0, \quad \forall k \geqq 0. \tag{3.14}
$$

Computing the determinants on the right-hand side of (3.13) and using (3.14), we find

$$
\begin{aligned}
F_\lambda(\varepsilon,0) &= -\varepsilon^4 (w_3'(\varepsilon,0,0)(w_4'''(\varepsilon,0,1) - w_4''(\varepsilon,0,1)) + w_4'''(\varepsilon,0,1) w_3(\varepsilon,0,1)) + O(\varepsilon^{-2}\delta_\varepsilon) \\
&= (q(0))^{-1/2} (q(1))^{-5/2} \sum_{0 \leqq k < N} c_k \varepsilon^k + O(\varepsilon^N)
\end{aligned} \tag{3.15}
$$

with any integer $N > 0$, where $c_0 = 1$ and c_k, $k > 0$, can be computed recursively using (2.8).

Thus, for each $\varepsilon \geqq 0$ fixed we have

$$
F(\varepsilon, \lambda) = F(\varepsilon, 0) + \lambda F_\lambda(\varepsilon, 0) + O(\lambda^2) \quad \text{as } \lambda \to 0 \tag{3.16}
$$

and for $|\lambda| \leqq \varepsilon^{-2}\delta_\varepsilon$ we find

$$
F(\varepsilon, \lambda) = F(\varepsilon, 0) + \lambda F_\lambda(\varepsilon, 0) + O(\varepsilon^{-4}\delta_\varepsilon^2). \tag{3.17}
$$

Hence, for the zero λ_0^ε of $F(\varepsilon, \lambda)$ in the interval $|\lambda| \leqq \varepsilon^{-2}\delta_\varepsilon$ we have the following asymptotic formula with any integer $N > 0$:

$$
\begin{aligned}
\lambda_0^\varepsilon &= -F(\varepsilon,0)(F_\lambda(\varepsilon,0))^{-1} + O(\varepsilon^{-4}\delta_\varepsilon^2) \\
&= 2(q(0))^{1/2}(q(1))^{-3/2}\delta_\varepsilon \left(\sum_{0 \leqq k < N} b_k \varepsilon^{k-1} + O(\varepsilon^N) \right), \qquad \varepsilon \to +0,
\end{aligned} \tag{3.18}
$$

where $b_0 = 1$, b_k, $k > 0$, can be computed recursively and where, of course, (3.11), (3.15), (3.17) have been used to derive (3.18).

Besides, for $\varepsilon \in (0, \varepsilon_0]$ with ε_0 sufficiently small, using (3.11) and (3.15), (3.17), we find

$$
F(\varepsilon, 0) < 0, \quad F(\varepsilon, \varepsilon) > 0 \quad \forall \varepsilon \in (0, \varepsilon_0],
$$

so that, in fact, $\lambda_0^\varepsilon \in (0, \varepsilon)$.

Thus, summarizing, we have proved the following theorem.

THEOREM 3.4. *The least eigenvalue* λ_0^ε *of* $\mathscr{A}^\varepsilon$ *defined by* (1.3)-(1.5) *is strictly positive for all* $\varepsilon \in (0, \varepsilon_0]$ *with* $\varepsilon_0 > 0$ *sufficiently small and, moreover, the following asymptotic formula holds for* λ_0^ε *with any integer* $N > 0$:

$$
\lambda_0^\varepsilon = 2(q(0))^{1/2}(q(1))^{-3/2}\delta_\varepsilon \left(\sum_{0 \leqq k < N} b_k \varepsilon^{k-1} + O(\varepsilon^N) \right) \quad \text{as } \varepsilon \to +0, \tag{3.19}
$$

where $b_0 = 1$, b_k, $k > 0$, can be computed recursively using (2.8) *and where δ_ε is given by* (3.12).

Next, we will prove the following theorem.

THEOREM 3.5. *For the eigenfunction $\psi_0^\varepsilon(x)$ of $\mathscr{A}^\varepsilon$ (given by* (1.3)-(1.5)) *associated with the least eigenvalue λ_0^ε the following asymptotic formula holds*:

$$\begin{aligned}\psi_0^\varepsilon(x) &= xC_1(\varepsilon) + \varepsilon C_2(\varepsilon) + \varepsilon C_3(\varepsilon) w_3(\varepsilon, 0, x) \\ &\quad + \lambda_0^\varepsilon((-x^3/6)C_1(\varepsilon) - (\varepsilon x^2/2)C_2(\varepsilon) + \varepsilon C_3(\varepsilon)\partial_\lambda w_3(\varepsilon, 0, x) - \varepsilon^2 w_4(\varepsilon, 0, x)) \\ &\quad + O((\lambda_0^\varepsilon)^2),\end{aligned} \tag{3.20}$$

where the coefficients $C_k(\varepsilon)$, $1 \leqq k \leqq 3$, are C^∞ functions in $\varepsilon \in [0, \varepsilon_0]$ admitting the asymptotic expansions:

$$C_k(\varepsilon) = \sum_{0 \leqq p < N} c_{kp}\varepsilon^p + O(\varepsilon^N), \tag{3.21}$$

with any integer $N > 0$, and for the functions $w_3(\varepsilon, 0, x)$, $\partial_\lambda w_3(\varepsilon, 0, x)$, $w_4(\varepsilon, 0, x)$ the asymptotic expansions (2.4) *with $u_{jk}(\lambda, x)$ defined by* (2.8) *are valid.*

Proof. We seek an eigenfunction $\psi_0^\varepsilon(x)$ of $\mathscr{A}^\varepsilon$ associated with the least eigenvalue λ_0^ε in the form

(3.22)

$$\psi_0^\varepsilon(x) = C_1(\varepsilon) w_1(\varepsilon, \lambda_0^\varepsilon, x) + \varepsilon C_2(\varepsilon) w_2(\varepsilon, \lambda_0^\varepsilon, x) + \varepsilon C_3(\varepsilon) w_3(\varepsilon, \lambda_0^\varepsilon, x) - \varepsilon^2 \lambda_0^\varepsilon w_4(\varepsilon, \lambda_0^\varepsilon, x),$$

where $w_k(\varepsilon, \lambda, x)$, $1 \leqq k \leqq 4$, is the fundamental system of solutions for (2.2) constructed in the proof of Theorem 2.3.

The boundary conditions for $\psi_0^\varepsilon(x)$ yield an overdetermined system for the coefficients $C_k(\varepsilon)$, $1 \leqq k \leqq 3$, which has nontrivial solutions, since $F(\varepsilon, \lambda_0^\varepsilon) = 0$ for all $\varepsilon \in (0, \varepsilon_0]$ with $F(\varepsilon, \lambda)$ defined by (3.2).

Thus, neglecting the equation for $C_k(\varepsilon)$, $1 \leqq k \leqq 3$, which results from the boundary condition $\partial_x^3 \psi_0^\varepsilon(\varepsilon, \lambda_0^\varepsilon, 1) = 0$, and rewriting the boundary conditions $\psi_0^\varepsilon(\varepsilon, \lambda_0^\varepsilon, 0) = 0$, $\partial_x^2 \psi_0^\varepsilon(\varepsilon, \lambda_0^\varepsilon, 1) = 0$ in the equivalent form $\varepsilon^{-1}\psi_0^\varepsilon(\varepsilon, \lambda_0^\varepsilon, 0) = 0$, $(\lambda_0^\varepsilon)^{-1}\partial_x \psi_0^\varepsilon(\varepsilon, \lambda_0^\varepsilon, 1) = 0$, for $C(\varepsilon) = (C_1(\varepsilon), C_2(\varepsilon), C_3(\varepsilon))$ we get the linear system

$$A(\varepsilon)C(\varepsilon) = g(\varepsilon), \tag{3.23}$$

where

$$A(\varepsilon) := \begin{Vmatrix} 0, & w_2(\varepsilon, \lambda_0^\varepsilon, 0), & w_3(\varepsilon, \lambda_0^\varepsilon, 0) \\ w_1'(\varepsilon, \lambda_0^\varepsilon, 0), & 0, & \varepsilon w_3'(\varepsilon, \lambda_0^\varepsilon, 0) \\ (\lambda_0^\varepsilon)^{-1} w_1''(\varepsilon, \lambda_0^\varepsilon, 1), & \varepsilon(\lambda_0^\varepsilon)^{-1} w_2''(\varepsilon, \lambda_0^\varepsilon, 1), & \varepsilon(\lambda_0^\varepsilon)^{-1} w_3''(\varepsilon, \lambda_0^\varepsilon, 1) \end{Vmatrix}, \tag{3.24}$$

$$g(\varepsilon) := (\varepsilon\lambda_0^\varepsilon w_4(\varepsilon, \lambda_0^\varepsilon, 0), \varepsilon^2 \lambda_0^\varepsilon w_4'(\varepsilon, \lambda_0^\varepsilon, 0), \varepsilon^2 w_4''(\varepsilon, \lambda_0^\varepsilon, 1)). \tag{3.25}$$

Indeed, as a consequence of Proposition 2.1, Lemma 2.2, and Theorem 2.3, we have

$$w_1(\varepsilon, \lambda, 0) = w_2'(\varepsilon, \lambda, 0) = 0.$$

Again using Proposition 2.1, Theorem 2.3, and (3.19), we easily find

$$A(\varepsilon) = A(0) + \varepsilon B(\varepsilon), \tag{3.26}$$

where

$$A(0) := \begin{Vmatrix} 0, & 1, & (q(0))^{1/2} \\ 1, & 0, & -(q(0))^{-1/2} \\ -1, & 0, & 2(q(0))^{-1/2} \end{Vmatrix}, \tag{3.27}$$

and where $B(\varepsilon)$ is a 3×3 matrix that is infinitely differentiable with respect to $\varepsilon\in[0,\varepsilon_0]$ and whose asymptotic expansion for $\varepsilon\to+0$ can be found explicitly:

$$B(\varepsilon)\sim\sum_{k\geqq 0}\varepsilon^k B_k,$$

using Proposition 2.1 and Theorem 2.3.

Since $A(0)$ is invertible, the matrix $A(\varepsilon)$ defined by (3.24) is invertible, as well, for all $\varepsilon\in[0,\varepsilon_0]$ with ε_0 sufficiently small and, moreover, $(A(\varepsilon))^{-1}$ is an infinitely differentiable matrix function of $\varepsilon\in[0,\varepsilon_0]$.

Furthermore, taking the first two terms in the Taylor expansion of $w_k(\varepsilon,\lambda,x)$, $1\leqq k\leqq 3$, with respect to λ around the point $\lambda=0$, and noticing that $w_1(\varepsilon,0,x)\equiv x$, $\partial_\lambda w_1(\varepsilon,0,x)=-x^3/6$, $w_2(\varepsilon,0,x)\equiv 1$, $\partial_\lambda w_2(\varepsilon,0,x)=-x^2/2$ (as a consequence of (2.5), (2.7), Lemma 2.2, and Theorem 2.3), we get asymptotic formula (3.20) with $C_k(\varepsilon)$, $1\leqq k\leqq 3$, admitting asymptotic expansion (3.21). □

Remark 3.6. An easy computation yields for the coefficients c_{kp} on the right-hand side of (3.21)

$$c_{10}=2(q(1))^{-3/2},\quad c_{20}=-2q(0)(q(1))^{-3/2},\quad c_{30}=2(q(0))^{1/2}(q(1))^{-3/2}. \tag{3.28}$$

Besides, as a consequence of (2.8), we also have

$$\begin{aligned} w_3(\varepsilon,0,x)&=((q(x))^{1/2}+O(\varepsilon))\exp\left(-\varepsilon^{-1}\int_0^x(q(y))^{-1}\,dy\right),\\ \partial_\lambda w_3(\varepsilon,0,x)&=O(\varepsilon)\exp\left(-\varepsilon^{-1}\int_0^x(q(y))^{-1}\,dy\right),\\ w_4(\varepsilon,0,x)&=((q(x))^{1/2}+O(\varepsilon))\exp\left(-\varepsilon^{-1}\int_x^1(q(y))^{-1}\,dy\right), \end{aligned} \tag{3.29}$$

so that, keeping only the main terms, we get the following asymptotic expansion for $\psi_0^\varepsilon(x)$:

$$\psi_0^\varepsilon(x)=2(q(1))^{-3/2}(1+O(\varepsilon))\left(x-\varepsilon q(0)+\varepsilon(q(0)q(x))^{1/2}\exp\left(-\varepsilon^{-1}\int_0^x(q(y))^{-1}\,dy\right)\right).$$

Moreover, the last formula may also be rewritten in the following equivalent form:

$$\psi_0^\varepsilon(x)=2(q(1))^{-3/2}(1+O(\varepsilon))(x-\varepsilon q(0)(1-\exp(-x/(\varepsilon q(0))))), \tag{3.30}$$

since freezing the coefficient $q(x)$ at $x=0$ brings over an error that is of order $O(\varepsilon^2\exp(-x/(\varepsilon q(0))))$.

Formula (3.30) can, of course, also be derived using the reduction method mentioned above in the Introduction.

Next, an asymptotic formula will be derived for an eigenfunction $\psi_n^\varepsilon(x)$ associated with an eigenvalue λ_n^ε, for which (2.7) is valid with a given integer $n>0$.

Seeking $\psi_n^\varepsilon(x)$ in the form

$$\begin{aligned} \psi_n^\varepsilon(x)&=\pi n w_1(\varepsilon,\lambda_n^\varepsilon,x)+\varepsilon C_2(\varepsilon)w_2(\varepsilon,\lambda_n^\varepsilon,x)\\ &\quad+\varepsilon C_3(\varepsilon)w_3(\varepsilon,\lambda_n^\varepsilon,x)+\varepsilon^3 C_4(\varepsilon)w_4(\varepsilon,\lambda_n^\varepsilon,x), \end{aligned} \tag{3.31}$$

the same argument as above in the case of $\psi_0^\varepsilon(x)$ yields a linear system for $C(\varepsilon)=(C_2(\varepsilon), C_3(\varepsilon), C_4(\varepsilon))$ with a 3×3 matrix $A(\varepsilon)$ that is infinitely differentiable with respect to $\varepsilon\in[0,\varepsilon_0]$ and such that

$$A(0)=\begin{Vmatrix} 1, & (q(0))^{1/2}, & 0 \\ 0, & -(q(0))^{-1/2}, & 0 \\ 0, & 0, & (q(1))^{-5/2} \end{Vmatrix}, \tag{3.32}$$

the second member $g(\varepsilon)$ of this linear system being

$$g(\varepsilon)=(0,-\pi n w_1'(\varepsilon,\lambda_n^\varepsilon,0),-\pi n w_1'''(\varepsilon,\lambda_n^\varepsilon,1)), \tag{3.33}$$

a C^∞ function of $\varepsilon\in[0,\varepsilon_0]$.

Therefore, the coefficients $C_k(\varepsilon)$, $2\leqq k\leqq 4$ on the right-hand side of (3.31) are well-defined C^∞-functions of $\varepsilon\in[0,\varepsilon_0]$, provided that ε_0 is sufficiently small. Besides, $C_k(\varepsilon)$ again admit asymptotic expansion:

$$C_k(\varepsilon)\sim\sum_{p\geqq 0} c_{kp}\varepsilon^p. \tag{3.34}$$

It is readily seen that

$$C_2(0)=-q(0)\pi n,\quad C_3(0)=(q(0))^{1/2}\pi n,\quad C_4(0)=(q(1))^{5/2}(-1)^n(\pi n)^3. \tag{3.35}$$

Thus, again freezing $q(x)$ at $x=0$ for w_3 and at $x=1$ for w_4, and using the fact that $\lambda_n^\varepsilon-\lambda_n^0=O(\varepsilon)$, from (3.31), (3.35) we get the following asymptotic formula for $\psi_n^\varepsilon(x)$:

$$\begin{aligned}\psi_n^\varepsilon(x)=\{&\sin\pi nx+\varepsilon q(0)(\exp(-x/(\varepsilon q(0)))-\cos\pi nx)\\ &+(-1)^n(\varepsilon q(1)\pi n)^3\exp(-(1-x)/(\varepsilon q(1)))\}(1+O(\varepsilon)).\end{aligned} \tag{3.36}$$

Also using (3.31) and Taylor's expansions in λ around $\lambda=\pi^2n^2$ of $w_k(\varepsilon,\lambda_n^\varepsilon,x)$, $1\leqq k\leqq 4$, for $\psi_n^\varepsilon(x)$ we obtain a full asymptotic expansion in the form

$$\begin{aligned}\psi_n^\varepsilon\sim&\sum_{p\geqq 0}\varepsilon^p u_{1p}(x)+\exp\left(-\varepsilon^{-1}\int_0^x(q(y))^{-1}\,dy\right)\sum_{p\geqq 1}\varepsilon^p u_{2p}(x)\\ &+\exp\left(-\varepsilon^{-1}\int_x^1(q(y))^{-1}\,dy\right)\sum_{p\geqq 3}\varepsilon^p u_{3p}(x),\end{aligned} \tag{3.37}$$

where, of course,

$$u_{10}(x)=\sin\pi nx,\qquad u_{11}(x)=-\varepsilon q(0)\cos\pi nx,$$
$$u_{21}(x)=(q(0)q(x))^{1/2},\qquad u_{33}(x)=(q(1))^{5/2}(q(x))^{1/2},$$

and all other coefficients $u_{pk}(x)$ can be computed recursively.

Remark 3.7. Of course, using the asymptotic formulae for $w_j(\varepsilon,\lambda,x)$, $1\leqq j\leqq 4$, in Theorem 3.1, we can establish asymptotic formulae for λ_0^ε and $\psi_0^\varepsilon(x)$ in (1.10) up to an error term $O(\varepsilon^\infty\delta_\varepsilon)$ with δ_ε given by (3.3), as $\varepsilon\to+0$.

For other eigenvalues λ_k^ε, $k=1,2,\cdots$, of (2.1) we may apply the reduction method developed in [6] and [7] as well, the procedure being very similar to the one in [4], since for the eigenvalues λ_k^ε, $k=1,2,\cdots$, of (1.10) we have

$$\lambda_k^\varepsilon=\pi^2k^2+O(\varepsilon),\qquad \varepsilon\to 0,$$

as was the case for $\mathscr{A}^\varepsilon$ in [4] (i.e., $\lambda_k^\varepsilon-\lambda_k^0$ for $k\geqq 1$ is no longer exponentially small, as $\varepsilon\to+0$, so that the parametrix used previously in the construction of a reducing operator for coercive singular perturbations $\mathscr{A}^\varepsilon$ may be used again with necessary minor modifications).

Remark 3.8. The following argument can be used to get heuristically the first term in the asymptotic expansion of the least eigenvalue λ_0^ε.

As a consequence of Theorem 2.3, we have the following asymptotic formulae for the fundamental system of solutions $w_k(\varepsilon, \lambda, x)$, $1 \leqq k \leqq 4$ of (2.2):

$$w_1(\varepsilon, \lambda, x) = \lambda^{-1/2} \sin(\lambda^{1/2}x) + O(\varepsilon^2),$$

$$w_2(\varepsilon, \lambda, x) = \cos(\lambda^{1/2}x) + O(\varepsilon^2),$$

$$w_3(\varepsilon, \lambda, x) = (1+O(\varepsilon))(q(x))^{1/2} \exp\left(-\varepsilon^{-1}\int_0^x (q(y))^{-1}\,dy\right),$$

$$w_4(\varepsilon, \lambda, x) = (1+O(\varepsilon))(q(x))^{1/2} \exp\left(-\varepsilon^{-1}\int_x^1 (q(y))^{-1}\,dy\right).$$

Using only the first terms in the asymptotic expansions for $w_k(\varepsilon, \lambda, x)$, $1 \leqq k \leqq 4$, and attempting to satisfy for a solution $u(\varepsilon, \lambda, x)$ of (2.2) the boundary conditions $u(\varepsilon, \lambda, 0) = \partial_x u(\varepsilon, \lambda, 0) = 0$, for $u(\varepsilon, \lambda, x)$ we get the asymptotic representation

$$u(\varepsilon, \lambda, x) \sim -\lambda^{-1/2} \sin(\lambda^{1/2}x) + \varepsilon q(0) \cos(\lambda^{1/2}x)$$
$$-\varepsilon(q(0)q(x))^{1/2}\left(\exp\left(-\varepsilon^{-1}\int_0^x (q(y))^{-1}\,dy\right)\right.$$
$$\left. + C_\varepsilon \exp\left(-\varepsilon^{-1}\int_x^1 (q(y))^{-1}\,dy\right)\right).$$

Trying to cancel the leading term in the asymptotic expansion for $\partial_x^3 u(\varepsilon, \lambda, 1)$, we conclude that the only reasonable choice for the constant C_ε on the right-hand side of the last formula is

$$C_\varepsilon = \delta_\varepsilon = \exp\left(-\varepsilon^{-1}\int_0^1 (q(y))^{-1}\,dy\right).$$

Furthermore, we also realize that the only way to satisfy asymptotically the boundary conditions $\partial_x^k u(\varepsilon, \lambda, 1) = 0$, $k = 2, 3$, is to have $\lambda = \lambda(\varepsilon) \to 0$ as $\varepsilon \to +0$.

Afterwards, replacing $\lambda^{-1/2} \sin(\lambda^{1/2}x)$ and $\cos(\lambda^{1/2}x)$ by $x - \lambda x^3/6$ and $1 - \lambda x^2/2$, respectively, and attempting to satisfy the boundary condition $\partial_x^2 u(\varepsilon, \lambda, 1) = 0$, we get in this heuristic way for λ_0^ε the first term in the asymptotic formula given by (3.19), and for the associated eigenfunction $\psi_0^\varepsilon(x)$ the first terms in the asymptotic formula given by (3.20).

The last term on the right-hand side of (3.20), which is of order $O(\delta_\varepsilon^2)$ at $x = 0$, seems to be redundant. However, it is not so, since at $x = 1$ this term is comparable with all other terms in the asymptotic formulae for $\partial_x^k \psi_0^\varepsilon(x)|_{x=1}$, $k = 2, 3, \cdots$. But away from the point $x = 1$, and especially in the neighbourhood of $x = 0$, only the first three terms on the right-hand side of (3.20) are relevant for the asymptotic behaviour of $\psi_0^\varepsilon(x)$ as $\varepsilon \to +0$. Of course, freezing $q(x)$ in the exponential term at the point $x = 0$, we get an asymptotic formula for $\psi_0^\varepsilon(x)$ valid in a neighbourhood of $x = 0$:

$$\psi_0^\varepsilon(x) = -x + \varepsilon q(0)(1 - \exp(-x/(\varepsilon q(0)))) + \varepsilon^2 w_\varepsilon,$$

with $\sup_\varepsilon \max_x |\partial_x^k w_\varepsilon(x)| < \infty$, $k = 0, 1, 2$, which may also be derived by using a reducing operator S^ε, constructed in the same way as in [6] and [7], i.e., by using the parametrix constructions where only the principal symbol of the coercive singular perturbation is involved.

Remark 3.9. It is readily seen that the reduced problem $\mathscr{A}^0 u^0 = (f, \varphi_1)^T$ is solvable if and only if the following condition is satisfied:

$$f(1) + \varphi_1(1) = 0,$$

the solution $u^0(x)$ being well defined up to the additive term Cx, which is the solution of the homogeneous problem $\mathscr{A}^0 x = (0, 0)^T$.

In other words, the index $\kappa(0)$ of the reduced problem is zero: $\kappa(0) = 0$.

Note that this is also true for the perturbed problem $\mathscr{A}^\varepsilon$, i.e., $\kappa(\varepsilon) = 0$, for all $\varepsilon \in (0, \varepsilon_0]$, since for all $\varepsilon \in (0, \varepsilon_0]$ with ε_0 sufficiently small the perturbed problem has a well-defined solution for any data $(f, \varphi_1, \varphi_2)^T$ (sufficiently smooth), i.e., dim ker $\mathscr{A}^\varepsilon$ = dim coker $\mathscr{A}^\varepsilon = 0$.

The stability of the index $\kappa(\varepsilon)$ of elliptic boundary value problems with respect to coercive singular perturbations is also a consequence of the general reduction procedure indicated in [6], [7], and [17].

4. Adjoint operator. We identify the singular perturbation $\mathscr{A}^\varepsilon$ given by (1.4) and (1.3) with the differential operator $L^\varepsilon = L(\varepsilon, 0, x, \partial_x) = \varepsilon^2 \partial_x^2 q^2(x) \partial_x^2 - \partial_x^2$ considered as an unbounded operator in $L_2(U)$ with the domain D_{L^ε},

$$D_{L^\varepsilon} := \{u \in H_4(U),\ \pi_{\partial U} B_k u(x') = 0,\ k = 1, 2\}, \tag{4.1}$$

where $H_4(U)$ is the Sobolev space of order 4.

Denote by $(L^\varepsilon)^*$ the adjoint of L^ε. The partial integration implies that $(L^\varepsilon)^*$ is the same differential operator $L(\varepsilon, 0, x, \partial_x)$ with the domain $D_{(L^\varepsilon)^*}$ defined as

$$D_{(L^\varepsilon)^*} := \{u \in H_4(U),\ \pi_{\partial U} {}^t B_k^\varepsilon(x', \partial_x) u(x') = 0,\ k = 1, 2\}, \tag{4.2}$$

where

$$\begin{aligned} {}^t B_1^\varepsilon(x, \partial_x) &= (x-1) + x(\varepsilon^2 q^2(x) \partial_x^2 - 1), \\ {}^t B_2^\varepsilon(x, \partial_x) &= (1-x) q^4(x) \partial_x q^{-2}(x) + x \partial_x (1 - \varepsilon^2 q^2(x) \partial_x^2). \end{aligned} \tag{4.3}$$

Hence, we may consider the adjoint operator $(L^\varepsilon)^* : D_{(L^\varepsilon)^*} \to L_2(U)$ of $L^\varepsilon : D_{L^\varepsilon} \to L_2(U)$ as a restriction (to homogeneous boundary conditions) of the operator $(\mathscr{A}^\varepsilon)^*$ defined as

$$(\mathscr{A}^\varepsilon)^* := (\pi_U r^\varepsilon D_x^2,\ \pi_{\partial U} {}^t B_1^\varepsilon,\ \pi_{\partial U} {}^t B_2^\varepsilon)^T. \tag{4.4}$$

Since the least eigenvalue λ_0^ε of $\mathscr{A}^\varepsilon$ defined by (1.4) is real (and strictly positive for $\varepsilon \in (0, \varepsilon_0]$), it coincides with the least eigenvalue of $(\mathscr{A}^\varepsilon)^*$ defined by (4.4), (4.3).

We will exhibit an asymptotic formula for the eigenfunction $\varphi_0^\varepsilon(x)$ of $(\mathscr{A}^\varepsilon)^*$ associated with the least eigenvalue $\lambda_0^\varepsilon > 0$.

Using Theorem 2.3 and the same argument as above for the eigenfunction ψ_0^ε of $\mathscr{A}^\varepsilon$ associated with λ_0^ε, we get the following asymptotic formula for $\varphi_0^\varepsilon(x)$:

$$\begin{aligned} \varphi_0^\varepsilon(x) \sim \delta_\varepsilon (2(q(0)q(1))^{1/2} x - 2(q(0))^{1/2} + (q(x))^{1/2} \exp\left(-\varepsilon^{-1} \int_0^x (q(y))^{-1}\, dy\right) \\ + (q(x))^{1/2} \exp\left(-\varepsilon^{-1} \int_x^1 (q(y))^{-1}\, dy\right). \end{aligned} \tag{4.5}$$

Neglecting terms that are $O(\delta_\varepsilon)$ uniformly with respect to $x \in \bar{U}$, we get the following simplified (and less accurate) asymptotic formula for the eigenfunction $\varphi_0^\varepsilon(x)$ of the adjoint problem:

$$\varphi_0^\varepsilon(x) \sim (q(x))^{1/2} \exp\left(-\varepsilon^{-1} \int_x^1 (q(y))^{-1}\, dy\right). \tag{4.6}$$

Furthermore, freezing $q(x)$ at the point $x=1$, for $\varphi_0^\varepsilon(x)$ we get an asymptotic formula that may be established using the reduction procedure from [6] and [7]:

$$\varphi_0^\varepsilon(x) \sim (q(1))^{1/2} \exp(-(1-x)/(\varepsilon q(1))). \tag{4.7}$$

For the normalized eigenfunction, i.e., φ_0^ε such that $|\langle \varphi_0^\varepsilon, \psi_0^\varepsilon \rangle| = 1$ with ψ_0^ε the corresponding eigenfunction of $\mathcal{A}^\varepsilon$ given by (3.2), (3.3), using (4.7) we find

$$\varphi_0^\varepsilon(x) \sim (\varepsilon q(1))^{-1} \exp(-(1-x)/(\varepsilon q(1))). \tag{4.8}$$

Note that the normalized eigenfunction converges to the Dirac's mass $\delta(1-x)$ at the point $x=1$.

The adjoint operator $(\mathcal{A}^\varepsilon)^*$ defined by (4.4) is not a coercive singular perturbation, the operator $\pi_U r^\varepsilon D_x^2$, however, still being an elliptic singular perturbation (see [3] for the definition of coercive singular perturbations).

We may wonder what the reduced operator $(\mathcal{A}^0)^*$, associated with $(\mathcal{A}^\varepsilon)^*$ given by (4.4), should be like. The answer is the formally defined reduced operator $(\mathcal{A}^0)^*$, where only the first boundary operator ${}^tB_1^\varepsilon(x, \partial_x)$ in (4.3) is kept and afterwards we set $\varepsilon = 0$ in both $\pi_U r^\varepsilon D_x^2$ and $\pi_{\partial U}{}^tB_1^\varepsilon$, the latter yielding

$$(\mathcal{A}^0)^* := (\pi_U D_x^2, \pi_{\partial U} 1)^T. \tag{4.9}$$

To realize that the formally defined operator (4.9) is indeed the reduced operator for $(\mathcal{A}^\varepsilon)^*$ given by (4.4), we must consider a suitable perturbation of $(\mathcal{A}^\varepsilon)^*$ by lower-order operators that shift the spectrum of $(\mathcal{A}^\varepsilon)^*$ away from zero. For instance, considering the boundary value problem (with $q(x) \equiv 1$ for simplicity)

$$\begin{aligned} &\pi_U(\varepsilon^2 D_x^2 + D_x^2 - \lambda) u_\lambda^\varepsilon = 0, \\ &\pi_{\partial U}{}^tB_k^\varepsilon u_\lambda^\varepsilon(x') = \varphi_k(x'), \qquad 1 \leqq k \leqq 2, \end{aligned} \tag{4.10}$$

with $\lambda \in \mathbf{C}$, $0 < |\lambda| < \pi^2$ and $\varphi_k(x')$, $k=1,2$ given and independent of ε, it is readily seen that there exists the pointwise limit

$$u_\lambda^0(x) = \lim_{\varepsilon \to 0} u_\lambda^\varepsilon(x) \quad \forall x \in [0,1) \tag{4.11}$$

and that, moreover, $u_\lambda^0(x)$ is the solution of the boundary value problem

$$\pi_U(D_x^2 - \lambda) u_\lambda^0(x) = 0, \qquad \pi_{\partial U} u_\lambda^0(x') = \varphi_1(x').$$

The same is true if we consider an inhomogeneous equation in (4.10) with a smooth second member $f(x)$ that does not depend on ε.

A somewhat surprising situation is the fact that the reduced operator $(\mathcal{A}^0)^*$ for $(\mathcal{A}^\varepsilon)^*$ defined by (4.4) no longer has $\lambda = 0$ as its eigenvalue, whereas for $\mathcal{A}^0$ defined by (1.6) and (1.3), zero is an eigenvalue with Cx as the associated eigenfunctions.

In fact, the solution $u_\lambda^\varepsilon(x)$ to (4.10) contains a singular part that converges (as $\varepsilon \to +0$) to $\gamma_\varphi \delta(1-x)$ with some constant γ_φ depending on $\varphi_1(x')$ and $\varphi_2(x')$, where $\delta(1-x)$ is the Dirac delta-function. An easy computation also shows that

$$\lim_{\varepsilon \to 0} \operatorname{Res} u_\lambda^\varepsilon(x)|_{\lambda=0} = \gamma_\varphi \delta(1-x) = \varphi_0^0(x)$$

so that, to some extent, it would be natural to consider $\delta(1-x)$ as an "eigenfunction" of $(\mathcal{A}^0)^*$.

Let us consider again the operator $(\mathcal{A}_\lambda^\varepsilon)^*$ associated with (4.10):

$$(\mathcal{A}_\lambda^\varepsilon)^* := (\pi_U(\varepsilon^2 D_x^4 + D_x^2 - \lambda), \pi_U{}^tB_1^\varepsilon, \pi_{\partial U}{}^tB_2^\varepsilon)^T, \qquad 0 < \lambda < \pi^2,$$

where of course ${}^tB_k^\varepsilon(x, \partial_x)$ are defined by (4.3) with $q(x) \equiv 1$, i.e.,

$$ {}^tB_1^\varepsilon(x, \partial_x) = (x-1) + x(\varepsilon^2 \partial_x^2 - 1), \qquad {}^tB_2^\varepsilon = (1-x)\partial_x + x\partial_x(1-\varepsilon^2\partial_x^2). $$

It is readily seen that the function

$$ u_\lambda^\varepsilon(x) = \varepsilon^{-1}(1+\varepsilon\lambda x/2)\exp(-(1-x)/\varepsilon), \qquad 0 < \lambda < \pi^2, $$

is the solution of the boundary value problem

$$ (\mathscr{A}_\lambda^\varepsilon)^* u_\lambda^\varepsilon(x) = (f^\varepsilon(x), \varphi_1^\varepsilon, \varphi_2^\varepsilon)^T, $$

with

$$ \begin{aligned} &f^\varepsilon(x) = (-\lambda^2 x/2)\exp(-(1-x)/\varepsilon), && \sup_{0<\varepsilon\leqq 1}\max_{x\in\bar{U}} |f^\varepsilon(x)| < \infty, \\ &\varphi_1^\varepsilon(x') = (1-x')\varepsilon^{-1}\exp(-\varepsilon^{-1}) + \varepsilon\lambda x', && \sup_{0<\varepsilon\leqq 1}\max_{x'\in\partial U} |\varphi_1^\varepsilon(x')| < \infty, \\ &\varphi_2^\varepsilon(x') = (1-x')\varepsilon^{-2}(1+\varepsilon\lambda/2)\exp(-\varepsilon^{-1}) + \lambda x', \\ & && \sup_{0<\varepsilon\leqq 1}\max_{x'\in\partial U} |\varphi_2^\varepsilon(x')| < \infty. \end{aligned} \tag{4.12} $$

Nevertheless,

$$ \max_{x\in\bar{U}} u_\lambda^\varepsilon(x) = \varepsilon^{-1}(1+\varepsilon\lambda/2) \to \infty \quad \text{as } \varepsilon \to +0. $$

Such a situation is impossible, for instance, for the following singular perturbation of $(\mathscr{A}_\lambda^0)^* = (\pi_U(D_x^2 - \lambda), \pi_{\partial U} 1)^T$:

$$ \mathscr{B}_\lambda^\varepsilon := (\pi_U(\varepsilon^2 D_x^4 + D_x^2 - \lambda), \pi_{\partial U} 1, \pi_{\partial U}\partial_n)^T, \qquad 0 \leqq \lambda < \pi^2, $$

which is coercive (see [3]), so that for the solution u_λ^ε of the problem

$$ \mathscr{B}_\lambda^\varepsilon u_\lambda^\varepsilon = (f^\varepsilon, \varphi_1^\varepsilon, \varphi_2^\varepsilon)^T $$

with the data $(f^\varepsilon, \varphi_1^\varepsilon, \varphi_2^\varepsilon)^T$ satisfying (4.12), we always have the following a priori estimate (a version of the maximum principle):

$$ \sup_{0<\varepsilon\leqq 1}\max_{x\in\bar{U}} |u_\lambda^\varepsilon(x)| \leqq C\left(\sup_{0<\varepsilon\leqq 1}\max_{x\in\bar{U}} |f^\varepsilon(x)| + \sum_{1\leqq k\leqq 2} \sup_{0<\varepsilon\leqq 1}\max_{x'\in\partial U} |\varphi_k^\varepsilon(x')| \right), $$

with some constant $C > 0$.

REFERENCES

[1] L. S. FRANK, *Problèmes aux limites coercifs avec un petit paramètre*, C.R. Acad. Sci. Paris Sér A, (1976), pp. 1109-1111.

[2] ———, *Perturbazioni singolari ellittiche*, Rend. Sem. Mat. Univ. Politec. Milano, 47 (1977), pp. 135-163.

[3] ———, *Coercive singular perturbations* I: *a priori estimates*, Ann. Mat. Pura Appl. (4), 19 (1976), pp. 41-113.

[4] ———, *Perturbations singulières coercives* IV: *problème des valeurs propres*, C.R. Acad. Sci. Paris Sér. I, 301, (1985), pp. 69-72.

[5] ———, *Perturbations singulières coercives*: *reduction à des perturbations régulières et applications*, Séminaire Equations aux Dérivées Partielles 1986-1987, exposé XVIII, Centre de Mathématiques, Ecole Polytechnique, Palaiseau, France, April 1987, pp. 1-26.

[6] L. S. FRANK AND W. D. WENDT, *Coercive singular perturbations* II: *reduction to regular perturbations and applications*, Comm. Partial Differential Equations, 7 (1982), pp. 469-535.

[7] ———, *Coercive singular perturbations* III: *Wiener-Hopf operators*, J. Analyse Math., 43 (1983/84), pp. 88-135.

[8] L. S. FRANK AND J. J. HEIJSTEK, *On the reduction of coercive singular perturbations*, in Proc. Conference on Operator Theory and Applications, Calgary, Alberta, Canada, 1988.
[9] T. KATO, *Perturbation theory of semi-bounded operators*, Math. Ann., 125 (1953), pp. 435-447.
[10] L. LANDAU AND E. LIFSCHITZ, *Théorie de l'elasticité*, MIR, Moscow, 1967 (translated from Russian).
[11] J. MOSER, *Singular perturbation of eigenvalue problems for linear differential equations of even order*, Comm. Pure Appl. Math., 8 (1955), pp. 251-278.
[12] LORD RAYLEIGH, *The Theory of Sound*, Vol. I, London, 1937.
[13] R. T. SEELEY, *Extension of C^∞ functions defined on a half-space*, Proc. Amer. Math. Soc., 15 (1964), pp. 625-626.
[14] M. I. VISHIK, *On strongly elliptic systems of differential equations*, Mat. Sb. (N.S.), 29 (1951), pp. 615-676.
[15] M. I. VISHIK AND L. A. LYUSTERNIK, *Regular degeneration and boundary layer for linear differential equations with small parameter*, Uspekhi Mat. Nauk, 12 (1957), pp. 3-122; Russian Math. Surveys, 20 (1962), pp. 239-364.
[16] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, Interscience, New York, 1965.
[17] W. D. WENDT, *Coercive singularly perturbed Wiener-Hopf operators and applications*, Ph.D. thesis, University of Nijmegen, Nijmegen, the Netherlands, 1983.

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1264-1280, September 1990
© 1990 Society for Industrial and Applied Mathematics
010

FREE LAYERS IN A SINGULARLY PERTURBED BOUNDARY VALUE PROBLEM*

ADRIANA BOHÉ†

Abstract. The jumps of solutions of the boundary value problem $\varepsilon x'' = g(x)f(x')$ $c < t < d$ $x(c) = a$, $x(d) = b$ are studied for a small, positive ε. Using continuity arguments, solutions with a free layer at t_0 are found for all $t_0 \in (c, d)$ and the position of t_0 as a function of ε and of the boundary conditions is given. A new phenomenon appears as a consequence of the sensitive dependence of t_0 on the boundary data. For certain values of a and b the location of t_0 changes rapidly on $[c, d]$. There is a sharp transition between solutions with boundary layers at c and those with boundary layers at d. The set of (a, b) for which there are free layers is given. Nonstandard analysis methods and the geometrical approach of the observability plane are used.

Key words. boundary value problem, singular perturbations, free layer, nonstandard analysis

AMS(MOS) subject classifications. 34B15, 34D15, 34E15, 26E35, 03H05

Introduction. We consider the singularly perturbed boundary value problem

$$\varepsilon x'' = g(x)f(x'), \qquad c < t < d, \tag{1}$$

$$x(c) = a, \quad x(d) = b, \quad a \not\equiv b, \tag{2}$$

where a and b are limited real numbers, ε is a fixed positive infinitesimal, and c and d are standard real numbers.

Problems such as (1), (2) arise in the study of compressible fluids in gas dynamics; cf., for example, [2], where ε is a parameter that is small when the viscosity is small.

This boundary value problem has solutions with interesting features like boundary layers and interior layers.

It is interesting, from both the physical and the mathematical point of view, to locate the interior jumps as a function of the boundary conditions.

This paper concerns the study of the jumps exhibited by the solutions of (1), (2) and the location of the position t_0 of these jumps, as well as determining the set of values of the boundary conditions a and b for which there are free layers. Our methods use nonstandard analysis[1] and the geometrical approach of the observability plane method [3].

Our main objective is to study a new phenomenon that appears as a consequence of the sensitive dependence of t_0 on the boundary conditions. On studying the position t_0 as a function of a and b we will prove by continuity arguments that for all $t_0 \in [c, d]$ there exists a pair of boundary conditions (a, b) such that the solution jumps at t_0.

In addition, we give a method for determining t_0 as a fuction of a and b and ε.

Then we study the effect of slight variations at one of the boundary conditions on the behavior of the solutions. For certain values of a and b, this perturbation rapidly changes the location of t_0 in $[c, d]$.

The transition between solutions with a jump at one endpoint and those with a jump at the other endpoint is continuous but very sharp. In fact, if a solution with a

* Received by the editors January 6, 1989; accepted for publication (in revised form) August 28, 1989.
† Université de Paris 7, U.F.R. de Mathématiques, 75251 Paris Cedex 05, France.

[1] We adopt in this article a nonstandard point of view, which turns out to be valuable when studying singular perturbation problems [8], [10], [13]. An appendix defining the nonstandard terminology used in this paper is included at the end of the paper.

free layer exists for the boundary conditions (a, b^*) when a is fixed, any other solution with a free layer may only be obtained for values of b in "a very near vicinity" of b^* ($b \simeq b^*$; i.e., in the halo of b^*) (see Fig. 1.1).

We characterize the values of (a, b) which ensure that t_0 corresponds to a free layer.

We can find classical studies referring to the existence of a solution of (1), (2) with a free layer and the location of t_0. O'Malley [11] has considered the f linear case. Chang and Howes [1] and Howes [6] have studied the f linear and f quadratic cases. There are also nonstandard studies for $f(x') = x'^{[s]}$, $0 < s \leqq 2$ by Diener [4], which only concern the behavior of the solution and its jump but not the location of the transition point.

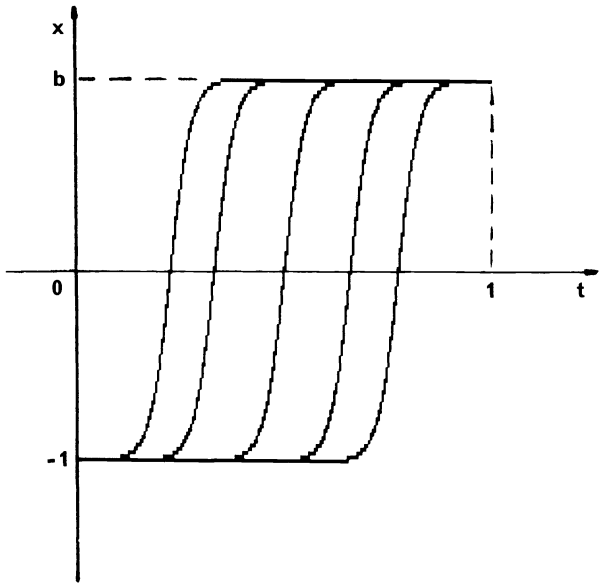


FIG. 1. *Solutions of* $\varepsilon x'' = -2xx'$, $x(0) = -1$, $x(1) = b$, *for five values of* $b \simeq 1$, *and* $\varepsilon = 0.05$.

When the boundary value problem is an autonomous problem like (1), (2) and the solutions of the reduced problem are constants, Chang and Howes [1] have shown that free layers are possible only when the boundary values satisfy a condition corresponding to the classical Rankine-Hugoniot condition in gas dynamics, and have found the location of the transition point in the quasilinear problem.

We determine the location of t_0 for a wide class of functions f, where f is not necessarily a power of x'. Also, in connection with the study of the flow of a compressible fluid, we find that a shock is possible not only when the Rankine-Hugoniot condition is satisfied but also when it is infinitely close to being satisfied.

The behavior of the transition point t_0 as a function of the boundary conditions is connected with another behavior that has been observed by Matkowsky [9] in the study of resonance in a quasilinear boundary value problem. Matkowsky has remarked that perturbations at one endpoint of the interval of $O(\varepsilon^\delta)$ for any δ change a solution with two boundary layers into a solution with only one boundary layer. Our results explain this phenomenon and make precise the order of the perturbation that causes significant changes in the solution.

To our knowledge, the phenomenon studied in this work has not been detected before. This is probably due to the fact that, classically, the behavior of the solutions

as ε tends to zero is studied for fixed boundary values. This obscures the dependence of the jump location on the boundary values. The nonstandard approach allows this dependence to appear very clearly.

Finally we note that we can expect this phenomenon to occur in a more general class of equations than (1), (2), since it may arise when a certain geometric situation (independent of the equation itself) takes place in the observability space, as we will see later.

In § 2 we give some nonstandard definitions of boundary layers and free layers, thickness and extremities of jumps, and we show briefly that problem (1), (2) has a unique, strictly monotone solution.

In § 3 we study the solution as a trajectory of the slow-fast system associated with (1). We prove that there is only one jump, and we study it in its observability plane. We give a geometrical picture of the behavior of the boundary and free layers.

We prove in § 4 that for every $t_0 \in [c, d]$ there exists a solution with a jump at $t = t_0$. We locate the transition point t_0 as a function of a and b and ε, and we give the values of (a, b) for which the solution has a free layer at t_0.

In § 5 we present some numerical results that illustrate this phenomenon in different examples.

Finally in § 6 we discuss other related examples, and a problem that arises in modelling compressible flows. In this case, we find a supersonic-subsonic shock not only when the boundary conditions satisfy the Prandtl shock relation, but also for values exponentially close to these conditions ($O(e^{-1/\varepsilon})$).

2. Preliminaries. We adopt the same definitions of a jump, its thickness, and its extremities as in [4].

DEFINITION 2.1. An internal function $x(t)$ has a *jump* in the interval $[t_1, t_2]$ if $x(t_1)$, $x(t_2)$ are limited values, $x(t_1) \not\approx x(t_2)$, and $x'(t)$ is unlimited on $[t_1, t_2]$. (A real number L is limited if its absolute value is smaller than some standard integer. A real number W is unlimited if it is larger than any standard integer.) We say that there is a *jump* at $t = t_0$ if there exists an interval $[t_1, t_2] \subset \text{hal}(t_0)$ in which there is a jump and t_0 is the *standard* part of t_1, t_2, i.e., t_0 is the unique standard that is infinitely close to t_1, t_2 (it is the standard $t_0 = {}^\circ t_1 = {}^\circ t_2$).

Remark 2.2. We note that any jump is monotone strictly increasing or decreasing since, by continuity, $x'(t)$ may not change its sign without becoming limited.

DEFINITION 2.3. We assume that $x(t)$ is a function with a jump in $[t_1, t_2]$ and that there is $T < t_1$ with $x(T)$ and $x'(T)$ limited, such that $x(t)$ is increasing for all $t \in [T, t_2]$. We say that the standard x_- is the origin of the jump in $[t_1, t_2]$ if for all $t \in [T, t_1]$ we have: If $x(t) \gg x_-$ $x'(t)$ is unlimited and if $x(t) \ll x_-$ there is $\tau \in [t, t_1]$ such that $x'(\tau)$ is limited.

The extremity x_+ of the jump is defined in the same way.

DEFINITION 2.4. Suppose $x : I \to \mathbb{R}$ has a jump in $[t_1, t_2] \subset I$ of extremities x_- and x_+. The thickness of the jump is the external set

$$\xi = \{t \in I : x_- \ll x(t) \ll x_+\}.$$

The thickness is a galaxy contained in hal (t_0), with $t_0 = {}^\circ t_1 = {}^\circ t_2$. (A set G is called a galaxy if G is external and if there is an internal sequence $(A_n)_{n \in \mathbb{N}}$ of internal sets such that $G = \bigcup_{n \in \mathbb{N}} A_n$.)

DEFINITION 2.5. We say that a solution of a boundary value problem has a *boundary layer* at c (or at d) if it has a *jump* at $t_0 = c$ (or at $t_0 = d$). We say that a solution of a boundary value problem has a *free layer* at t_0 if the solution has a *jump* at $t_0 \in (c, d)$.

We assume that f and g are internal functions that satisfy the following hypothesis H:

(i) $g \in C(\mathbb{R})$, g locally Lipschitz and f, g are of class S^0,

(ii) $f \in C^1(\mathbb{R})$; $f(0)=0$; $f(\nu)>0$ if $\nu>0$ and $f(\nu)<0$ if $\nu<0$,

(iii) $$\int_{\nu_0}^{+\infty} \frac{\nu}{f(\nu)}\, d\nu = +\infty,$$

where ν_0 is a limited positive constant.

f is the *type of growth* of (1) for unlimited ν:

(iv) The function $G:\mathbb{R}\to\mathbb{R}$, $G(x)=\int_a^x g(u)\,du$ satisfies $b^* \gg a$ ($b^* \ll a$) such that

$$G(b^*)=0, \quad G'(a)\gg 0, \quad G'(b^*)\ll 0,$$

$G(x)\gg 0$ for all $a \ll x \ll b^*$ ($G(x)\ll 0$ for all $b^* \ll x \ll a$).

We assume in condition (ii) that the appropriate reduced problem has only constant solutions. Condition (iii), which implies, in particular, the so-called Nagumo condition, restricts us to functions f having at most a quadratic growth.

Finally, in (iv) we assume, when a is fixed, the existence of the value b^* that satisfies $G(b^*)=0$, which is nothing more than the Rankine–Hugoniot shock condition.

We are interested in the behavior of the solutions of (1), (2) for b infinitely close to b^*.

The existence of a solution $x(t)$ of (1), (2), such that $a \leqq x(t) \leqq b$, follows from Jackson's version of the classic theorem of Nagumo [7, p. 354], with $\alpha \equiv a$ and $\beta \equiv b$.

Monotonicity is ensured by the following lemma.

LEMMA 2.6. *Suppose $f \in C^1(\mathbb{R})$, $g \in C(\mathbb{R})$, g locally Lipschitz. Then the solution $x(t)$ of* (1), (2) *is strictly monotone.*

Proof. As a consequence of the uniqueness of solution of the initial value problem associated with (1), the solutions of (1), (2) are strictly monotone. In fact, if we suppose that $x(t)$ has a maximum (or a minimum) M at $\tau \in (c, d)$, then the initial value problem

$$\varepsilon x'' = g(x)f(x'),$$

$$x(\tau)=M, \qquad x'(\tau)=0$$

would have two different solutions $x(t)$ and $x(t)\equiv M$.

From Lemma 2.6 we deduce that the inverse function $t(x)$ of $x(t)$ exists, is also strictly monotone in $[a, b]$, and satisfies the next boundary value problem:

$$\varepsilon t'' = -g(x)f(1/t')t'^3, \qquad a<x<b, \tag{3}$$

$$t(a)=c, \qquad t(b)=d. \tag{4}$$

From the application of the maximum principle [12] to (3), (4), it turns out that the solution $t(x)$ is unique. This allows us to deduce the uniqueness of the solution $x(t)$ of (1), (2).

The relationship between $x(t)$ and its inverse $t(x)$ plays an important part in the study and description of the behaviors of the solution to problem (1), (2).

Remark 2.7. A direct application of the maximum principle to boundary value problem (1), (2) requires $g(x)$ to be such that $G(x)$ results a convex function in $[a, b^*]$ for $x'>0$, or a concave function in $[b^*, a]$ for $x'<0$. Such restriction on the sign of $G''(x)$ is not possible under the hypothesis given above when we are interested in free layers.

3. Slow-fast system.

3.1. To describe the solutions of (1), (2) for a fixed infinitesimal ε, let us consider the differential equation (1) as the slow-fast system in the phase space (t, x, ν):

$$\text{(5)} \qquad \begin{cases} t' = 1, \\ x' = \nu, \\ \varepsilon\nu' = g(x)f(\nu), \end{cases}$$

where $\gamma(t) = (t, x(t), \nu(t))$ is the trajectory associated to a solution $x(t)$ of (1). A solution of the boundary value problem (1), (2) is, then, an integral curve $\gamma(t)$ of (5), which starts at the vertical line r_a defined by $t = c$, $x = a$ and reaches r_b defined by $t = d$, $x = b$ (see Fig. 3.1).

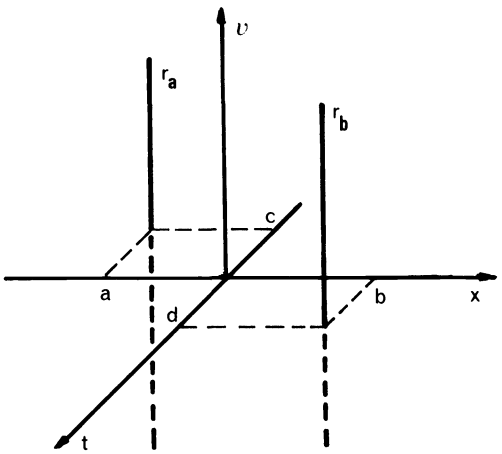


FIG. 3.1

From the properties of slow-fast systems it turns out that the solution of (1), (2) must necessarily jump in order to satisfy the other boundary condition. If not, as the slow portions of $\gamma(t)$ (for $\nu = x'$ limited) are in the halo of the slow surface S defined by the horizontal plane $\nu = 0$ and the vertical plane $x = {}^\circ\alpha$, where ${}^\circ g({}^\circ\alpha) = 0$, the solution would be almost constant in the whole interval $[c, d]$ with $a \not\simeq b$, which is absurd (see Fig. 3.2).

In order to study the trajectories of (1) with rapid motions (for $\nu = x'$ unlimited), which are outside the limited region (called main galaxy $\mathbb{G}$) of the space (t, x, ν), we use the observability space (t, x, V) [3].

We will use the transformation given by $\nu = h(V/\varepsilon)$, where h is the diffeomorphism of class S^0, $h : [0, +\infty) \to [\nu_0, +\infty)$ defined by

$$\text{(6)} \qquad hh' = f(h), \qquad h(0) = \nu_0.$$

The rapid motions of $\gamma(t)$ related to the jump at t_0 are, in the observability space, contained in the halo of the vertical plane of equation $t = t_0$ (called the observability plane) and are infinitely close to the curves (see Fig. 3.3):

$$V(x(t)) \simeq G(x) + C, \qquad t \simeq t_0.$$

The slow motions of $\gamma(t)$ appear in this space, in the halo of $V = 0$.

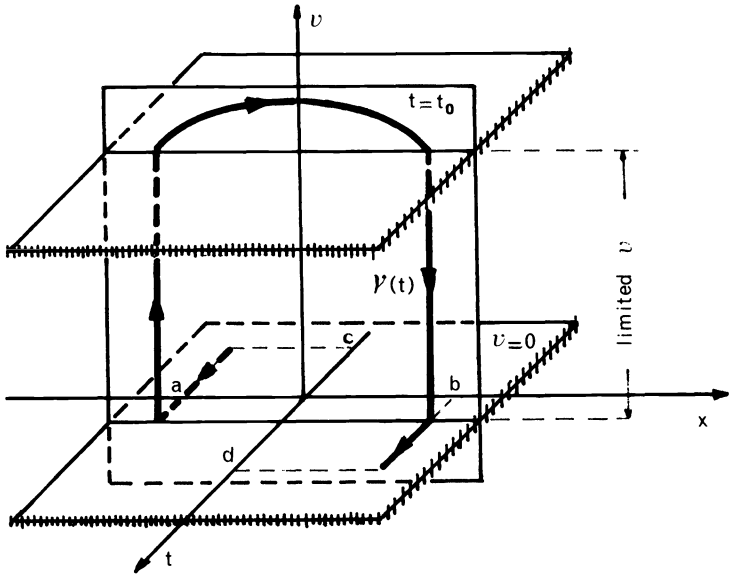


FIG. 3.2. *Trajectory $\gamma(t)$ of (5) associated with the solution $x(t)$ with a jump at $t=t_0$ in the phase space.*

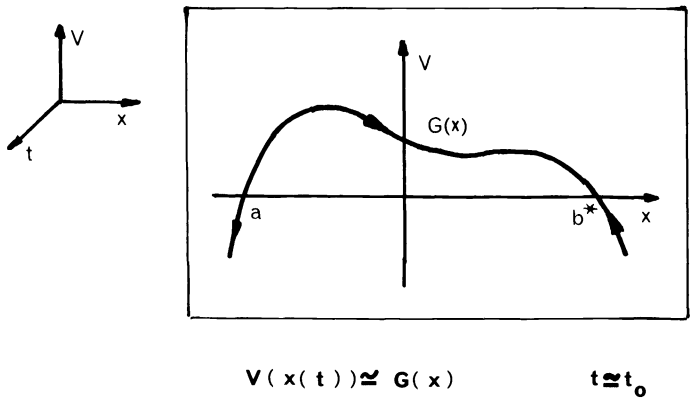


FIG. 3.3. *The trajectory of (5) associated with the solution of (1), (2) for $b=b^*$, in the observability plane $t=t_0$.*

The rapid motions of any trajectory related to a jump at t_0 move, in the observability plane $t=t_0$, from the left to the right in the positive half-plane (and from the right to the left in the negative portion). In Fig. 3.3 we have drawn only one curve, which represents the rapid portion of a trajectory associated with the solution of the boundary value problem for $b=b^*$ in the case of an increasing jump ($V>0$). In the negative region $V<0$, the two portions of trajectories correspond to the rapid motion of trajectories with decreasing jumps. The arrows on the curve $V=G(x)$ (Fig. 3.3) show the direction in which the point $(x, V(x))$ describes each trajectory.

Remark 3.1.1. When $\gamma(t)$ reaches the halo of $V=0$ in $x=x_0$ with $V(x_0)=0$, $V'(x_0)\simeq G'(x_0)\neq 0$, Theorem 3 [4, p. 552] ensures that $(^\circ x_0)$ is the extremity of the jump, i.e., $\gamma(t)$ has finished jumping. Then it remains in the halo of $V=0$ and it continues as a slow trajectory. Therefore condition (iv) in hypothesis H ensures that the standard part of a and b^* are the origin and the extremity of the increasing jump (see Fig. 3.3).

LEMMA 3.1.2. *Suppose f and g satisfy hypothesis* H; *then for any $a \ll b \leqq b^*$ and for any $b \in \text{hal}(b^*)$, the solution $x(t)$ of* (1), (2) *has only one jump.*

Proof. As the solution $x(t)$ is strictly increasing and near constant in its slow portions, then if we suppose there are two jumps, there exists $a \ll \beta \ll b$ such that $x(t) \simeq \beta$ with $x'(t) \simeq 0$. Then the trajectory associated with this slow portion must reach, in the plane (x, V), the halo of $V = 0$, i.e., $G(\beta) \simeq 0$. But this is not possible because, during each jump, the trajectory of (5) is such that $V(x) \simeq G(x) \gg 0$ for $a \ll x \ll b^*$.

3.2. Boundary and free layers. Geometrical features. Let us now investigate the existence of boundary layers and free layers by studying the related trajectory $\Gamma(t) = (t, x(t), V(x(t)))$ in the observability space.

Let us assume that a is fixed, and consider the surface S^* defined by

$$S^* = \{(t, x, V): t \in [c, d],\ x \in [a, b^*],\ V = G(x)\}.$$

S^* is built up by considering any possible jump in $[c, d]$ from the slow solution $x = a$ (see Fig. 3.4).

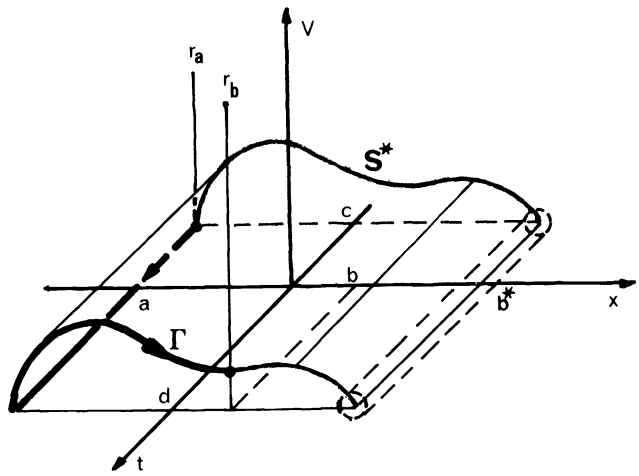


FIG. 3.4. *The trajectory $\Gamma(t)$, associated with the solution $x(t)$ with a boundary layer at d for $b \ll b^*$, in the space (t, x, V) and the surface S^* of the jumps.*

Note 3.2.1. Since $G(x)$ is a function of class S^1 and b^* is such that $G(b^*) = 0$ and $G'(b^*) = g(b^*) \ll 0$, then there exists a standard $s \gg b^*$, such that $G(b) \ll 0$ for $b^* \ll b \leqq s$.

LEMMA 3.2.2. *Suppose hypothesis* H *is satisfied, then for any b:*

(i) *$a \ll b \ll b^*$ there is a boundary layer at $t = d$,*

(ii) *$b^* \ll b \leqq s$ there is a boundary layer at $t = c$.*

Proof. (i) Since $a \ll b \ll b^*$, $V(b) \gg 0$, then, in order to join r_a with r_b, $\Gamma(t)$ must remain in the halo of $V = 0$ with $x(t) \simeq a$ and must reach r_b with a jump in the halo of $t_0 = d$. If not, if we suppose that $x(t)$ jumps at $t_0 \neq d$, $\Gamma(t)$ would reach the halo of $V = 0$ with $x(t) \simeq b^* \gg b$ and $V'(b^*) \simeq g(b^*) \neq 0$. By virtue of Remark 3.1.1, $^\circ(b^*)$ would be the extremity of the jump, and as there is only one jump (Lemma 3.1.2), $\Gamma(t)$ would remain in the halo of b^*. Then $\Gamma(t)$ would never be able to reach r_b, which is absurd because there is a solution.

(ii) Since $b^* \ll b \leqq s$, $G(b) \ll 0$, then the curve $V(x) = G(x) - G(b)$ is such that $V(a) = -G(b) \gg 0$. In order to join r_a with r_b, $\Gamma(t)$ must jump at $t = c$, i.e., it must leave r_a at $V(a)$ to reach the halo of $V = 0$ with $x(t) \simeq b$. If we suppose the contrary, arguing as in (i), we arrive at an absurdity.

On the contrary, if $b \simeq b^*$, $V(b) \simeq 0$, then any $t_0 \in [c, d]$ will allow $\Gamma(t)$ to join r_a with r_b.

The following proposition states that we can also find, in the halo of b^*, solutions with boundary layer at one endpoint or at the other.

PROPOSITION 3.2.3. *Assume hypothesis* H *is satisfied; then there exist* b_1, b_2 *in the halo of* b^* *such that the solution* $x(t)$ *of* (1), (2) *has a boundary layer at* $t_0 = c$ *if* $b = b_2$ *or at* $t_0 = d$ *if* $b = b_1$.

Proof. Let us consider the following external sets:

$$H_1 = \{b: a \ll b < b^* \text{ and } x'(d) \simeq +\infty\}, \qquad G_1 = \{b: a \ll b \ll b^*\}.$$

H_1 is a halo and G_1 is a galaxy such that $G_1 \subset H_1$.

It follows from Fehrele's Principle that $G_1 \subsetneqq H_1$. (The Fehrele Principle is a permanence principle stating that a halo cannot be at the same time a galaxy; it shows the incompatibility of certain kind of set.) Thus there is $b_1 \leqq b^*$ for which the solution of (1), (2) with $b = b_1$ has a boundary layer at $t = d$.

In the same way we prove that there is $b_2 \geqq b^*$ for which the solution has a boundary layer at $t = c$ if we consider the halo H_2 and the galaxy G_2: $H_2 = \{b: b^* < b \leqq s$ and $x'(c) \simeq +\infty\}$, $G_2 = \{b: b^* \ll b \leqq s\}$.

4. Continuous dependence of the location of the jump on b.

4.1. Continuity and S-continuity of t_0. Let us fix a. For each b there is a unique m_a such that the solution $x(t)$ of (1), (2) satisfies $x'(c) = m_a > 0$, as a consequence of the uniqueness of solutions to the boundary value problem.

Let us consider the family of initial value problems

$$\varepsilon x'' = g(x) f(x'), \tag{7}$$

$$x(c) = a, \qquad x'(c) = m_a. \tag{8}$$

The inverse function $t(x)$ of each solution of (7), (8), is a solution of the parameter family of problems

$$t' = H(x, m_a), \tag{9}$$

$$t(a) = c, \tag{10}$$

where $H(x, m_a) = 1/(h(G(x)/\varepsilon + h^{-1}(m_a)))$ is of class C^1 with respect to x and m_a, and h is the diffeomorphism defined by (6).

It follows from the theorem of continuous dependence of solutions on parameters that $t(x, m_a)$, the solution of (9), (10), is of class C^1 with respect to x and m_a. In addition, b and m_a satisfy the implicit relation $t(b, m_a) = d$, where $\partial t/\partial m_a(b, m_a) \neq 0$ for $m_a > 0$, $b > a$.

Then, as a consequence of the Implicit Function Theorem, there exists a neighborhood U_0 of b and a unique function $\psi(b) \in C^1$, such that $\psi(b) = m_a$, $t(b, \psi(b)) = d$, for all $b \in U_0$.

Now one of the principal results of this section is the following theorem.

THEOREM 4.1.1. *For any standard* $t_0 \in [c, d]$, *there exist* a *and* b *such that the solution* $x(t)$ *of* (1), (2) *has a jump at* t_0.

Proof. Let us consider $a \ll b_1 \ll b^*$ and $b_2 \geqq b^*$ the values given by Proposition 3.2.3.

We define the function $\phi:[b_1, b_2]\to[c, d]$ by

$$\phi(b) = t((a+b)/2, \psi(b)) = c + \int_a^{(a+b)/2} \frac{1}{H(x, \psi(b))}\, dx,$$

i.e., $\phi(b)=\tau$, such that the solution $x_b(t)$ of (1), (2) satisfies $x_b(\tau)=(a+b)/2$. ϕ is well defined because of the uniqueness of solution of (1), (2) and $\phi(b_1)=\tau_1$, ${}^\circ\tau_1 = d$, $\phi(b_2)=\tau_2$, ${}^\circ\tau_2 = c$. Finally, ϕ is a continuous function in $[b_1, b_2]$. Therefore, for any $\tau\in[\tau_1, \tau_2]$, there exists $b\in[b_1, b_2]$ such that $\phi(b)=\tau$. Then, for any standard $t_0\in[c, d]$, there exists b such that the solution $x_b(t)$ has a jump at $t_0 = {}^\circ\tau$.

Remark 4.1.2. ϕ is an internal function of class C^1. As it turns out from Proposition 3.2.3, ϕ is not S-continuous at $b=b^*$ ($b_1\simeq b_2$ but $\phi(b_1)\not\simeq\phi(b_2)$).

4.2. The position of t_0. If $x(t)$ has a free layer at t_0, then $t(x)$ has two boundary layers at $x={}^\circ(a)$ and at $x={}^\circ(b)$. If $x(t)$ has a boundary layer at c, $t(x)$ has a boundary layer at ${}^\circ b$ (see Figs. 4.1 and 4.2).

The observability plane of the jump of $t(x)$ is defined by the diffeomorphism $\hat{h}$:

$$\hat{h} = \hat{h}^2 f(1/\hat{h}), \quad dt/dx = \hat{h}(\hat{V}/\varepsilon), \quad \hat{h}(0) = \hat{v}_0. \tag{11}$$

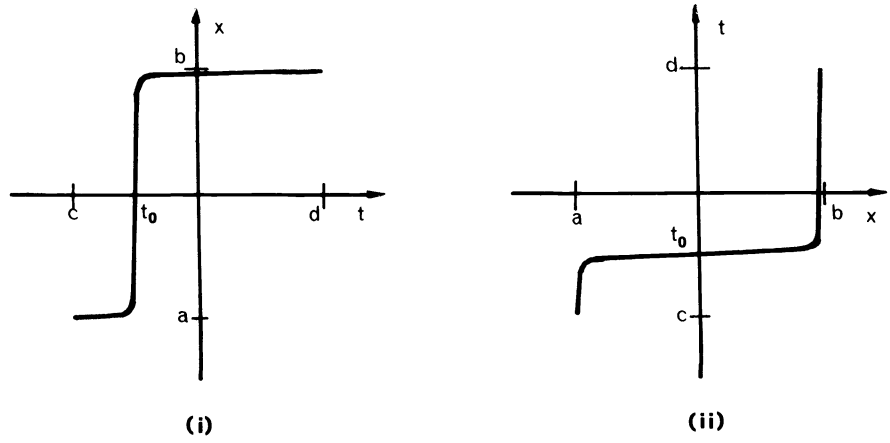


FIG 4.1. (i) *Solution of* (1), (2) *with a free layer.* (ii) *Corresponding solution of* (3), (4) *with two boundary layers.*

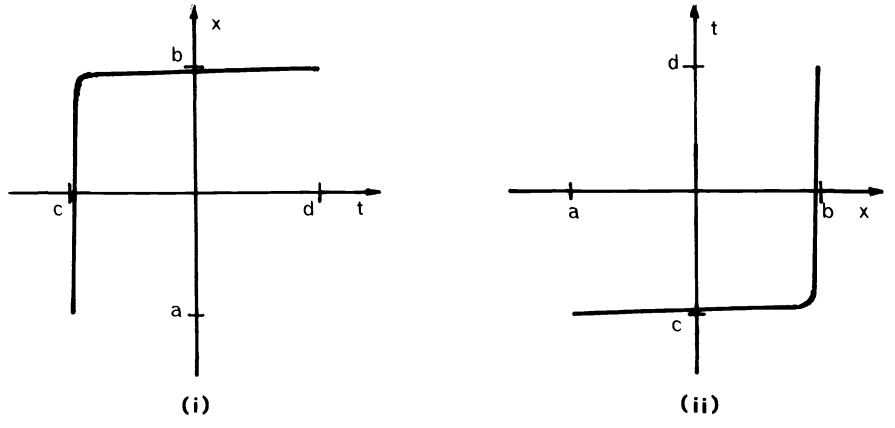


FIG. 4.2. (i) *Solution of* (1), (2) *with a boundary layer.* (ii) *Corresponding solution of* (3), (4) *with a boundary layer.*

Each jump of $t(x)$ is, in $(t, \hat{V})$, infinitely close to straight lines of slope $(-g(a))$ or $(-g(b))$, and we can describe the behavior of the trajectory associated with $t(x)$ as follows (see Fig. 4.3).

PROPOSITION 4.2.1. *Problem* (1), (2) *has a free layer for* $b = b^*$.

Proof. If we suppose there exists a boundary layer at c, the inverse function $t(x)$ of the solution $x(t)$ has a boundary layer at b^*. This jump is, in its observability plane $(t, \hat{V})$, infinitely close to

$$\hat{V} \simeq -g(b^*)(t-d) + \hat{V}(d).$$

Since $G(b^*) = 0$, $x'(c) = x'(d)$, then $\hat{V}(d) = \hat{V}(c)$, and since $t(x) \simeq c$ for $x \simeq a$, $\hat{V}(c) \simeq 0$. Then $g(b^*)(c-d) \simeq 0$ with $g(b^*) \neq 0$ and $c \neq d$, which is absurd. In a similar way we prove that a boundary layer at d is not possible.

PROPOSITION 4.2.2. *If problem* (1), (2) *has a free layer at* t_0, *then*

$$t_0 \simeq \frac{g(a)c - g(b)d}{g(a)-g(b)} - \frac{\varepsilon}{g(a)-g(b)} \left\{ \hat{h}^{-1}\left(\frac{1}{h((G(b)/\varepsilon) + h^{-1}(\psi(b)))} \right) - \hat{h}^{-1}(1/\psi(b)) \right\}, \tag{12}$$

where $\psi(b)$ *is defined implicitly by*

$$d - c = \int_a^b \frac{dx}{h(G(x)/\varepsilon + h^{-1}(\psi(b)))}.$$

Proof. As $x(t)$ has a free layer, $t(x)$ has two boundary layers, each of which is in the plane $(t, \hat{V})$ close to

$$\hat{V}_a(t) \simeq -g(a)(t-c) + \hat{V}(c), \qquad x \simeq a,$$
$$\hat{V}_b(t) \simeq -g(b)(t-d) + \hat{V}(d), \qquad x \simeq b.$$

As $x'(d) = h(h^{-1}(x'(c)) + G(b)/\varepsilon)$ and $t'(b) = \hat{h}(\hat{V}(d)/\varepsilon)$ and $x'(c) = \psi(b)$, we have that $\hat{V}(d) = \varepsilon \hat{h}^{-1}(1/(h(G(b)/\varepsilon + h^{-1}(\psi(b)))))$, $\hat{V}(c) = \varepsilon \hat{h}^{-1}(1/\psi(b))$. Then, as $\hat{V}_a(t_0) \simeq \hat{V}_b(t_0)$, equation (12) follows easily.

Remark 4.2.3. Proposition 4.2.2 gives an estimate of the transition point t_0 for problem (1), (2). The second term of this estimate is zero when $b = b^*$ since $G(b^*) = 0$ and, in this case,

$$t_0^* \simeq \frac{g(a)c - g(b^*)d}{g(a) - g(b^*)}.$$

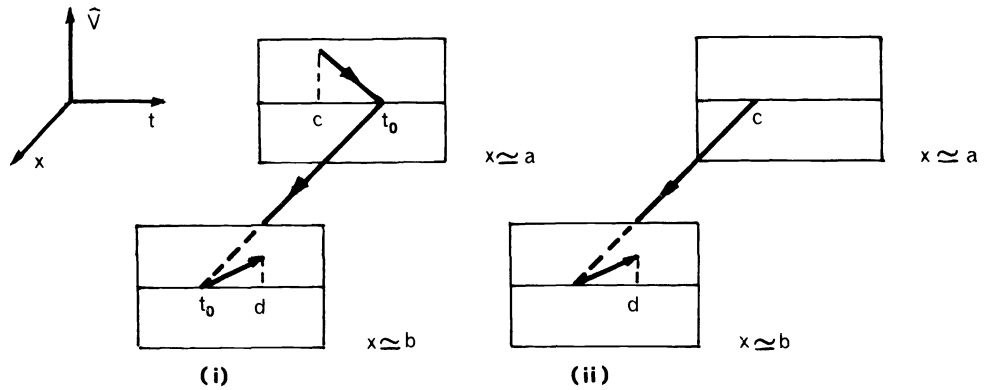


FIG. 4.3. *Trajectory corresponding to a solution of* (3), (4) *with* (i) *two boundary layers*; (ii) *one boundary layer.*

Chang and Howes [1] have found the same location of t_0^* for the quasilinear problem: $\varepsilon x'' = g(x)x'$, $c < t < d$, $x(c) = a$, $x(d) = b$, considering only $b = b^*$. However, for certain values of $b \neq b^*$ but in the halo of b^*, the second term of (12) is not infinitesimal and the transition is then located at $t_0 \neq t_0^*$.

4.3. The galaxy $\hat{\xi}$ of the free layer. Let us consider the function ϕ, which has been defined in Theorem 4.1.1. ϕ is a monotone, decreasing function, having a jump at $b_0 = {}^\circ(b^*)$, of extremities c and d.

The thickness of the jump of a solution to a singular perturbed second-order problem is known for a wide class of equations (see [4].)

In this case, the thickness of the jump of ϕ is given by the following galaxy $\hat{\xi}$:

$$\hat{\xi} = \{b\colon c \ll \phi(b) \ll d\}.$$

From Proposition 4.2.1 we know that $b^* \in \hat{\xi}$. We will see that $\hat{\xi}$ characterizes the set of values of b for which there is a free layer, as the next theorem shows.

THEOREM 4.3.1. *Problem* (1), (2) *has a free layer in* $[c, d]$ *if and only if* $b \in \hat{\xi}$.

Proof. ($\Rightarrow$) If $x(t)$ is a solution with a free layer, $\phi(b) = \tau$ is such that $t_0 = {}^\circ(\tau) \in (c, d)$, and thus $b \in \hat{\xi}$.

($\Leftarrow$) Let us consider $b \in \hat{\xi}$. Then $\tau = \phi(b)$ is such that the solution $x(t)$ satisfies $a \ll x(\tau) \ll b$ since $x(\tau) = (a+b)/2$. Then τ belongs to the thickness galaxy ξ of the jump of $x(t)$ since the extremities of this jump are ${}^\circ(a)$ and ${}^\circ(b)$. Therefore $x(t)$ jumps at $t_0 = {}^\circ(\tau)$ with $t_0 \in (c, d)$ as $b \in \hat{\xi}$. Then $x(t)$ has a free layer.

The thickness galaxy of a jump depends only on the type of growth of the equation with respect to x' unlimited. As ϕ is a solution of a singular perturbed problem with a type of growth $f(1/\phi')\phi'^3$ for ϕ' unlimited, the galaxy $\hat{\xi}$ is given by

$$\hat{\xi} = \left\{ b\colon b = b^* + \int_{\phi(b^*)}^{\phi(b)} \frac{d\phi}{\hat{h}(\hat{V}/\varepsilon)} \right\}.$$

Its determination is not difficult in many examples that have been considered in classical literature. For example, if f is linear, $\hat{\xi} = \{b \in \mathbb{R}\colon |b - b^*| < e^{-1/L^2\varepsilon},\ L \in \mathbb{G}\}$, which is called the ε-microgalaxy (b^*). If f is a power function of x' with $1 < s \leqq 2$, then $\hat{\xi} = \{b \in \mathbb{R}\colon b - b^* = \varepsilon^{3-s}L,\ L \in \mathbb{G}\}$ is called the ε^{3-s}-galaxy (b^*).

Remark 4.3.2. It follows from Theorem 4.3.1 that the transition of a boundary layer at one endpoint to a boundary layer at the other endpoint, passing along the free layers, is a phenomenon that may be observed only in the very little galaxy $\hat{\xi}$. That explains why this behavior has appeared, in general, as a discontinuity. Notice that, in the simple linear case $f(x') = x'$, this behavior is much sharper than in the others, as it takes place in the ε-microgal (b^*)—that is, for values of b that are exponentially close to b^*, instead of b such that $b - b^* = O(\varepsilon^{3-s})$ if $f(x') = x'^s$.

Remark 4.3.3. We note that, in general, this phenomenon would take place every time, in the observability space, a surface built up from a repelling slow solution by considering any possible jump in $[c, d]$ falls in with another one, built up from an attracting slow solution in the whole interval $[c, d]$.

Then let us consider a more general class of problem:

$$\varepsilon x'' = f(t, x, x'), \tag{13}$$

$$x(c) = a, \qquad x(d) = b, \tag{14}$$

where we assume that $f(t, x, x') = g(t, x)F(x') + r(t, x, x')$, where $F(x')$ is the type of growth of f, such that $r(t, x, x')/F(x') \simeq 0$ for x' unlimited.

Suppose that $u_1(t)$ is a solution of $°f(t, u, u') = 0$, $u(c) = a$, and $(\partial °f/\partial x')(t, u_1, u_1') > 0$, and $u_2(t)$ is a solution of $°f(t, u, u') = 0$, $u(d) = b$, and $(\partial °f/\partial x')(t, u_2, u_2') < 0$.

Then this phenomenon would occur in a problem such as (13), (14) provided that

$$\int_{u_1(t)}^{u_2(t)} g(t, x)\, dx \equiv 0 \quad \text{for all } t \in [c, d].$$

Remark 4.3.4. Finally we remark that our results explain the "strange phenomena" reported by Matkowsky [9] in a quasilinear boundary value problem. Matkowsky has noted that the solution $t(x)$ of the problem

$$\varepsilon t'' = 2\lambda(\varepsilon) x t' - 2\lambda(\varepsilon) n t, \qquad n = 0, 1, 2, \cdots, \tag{15}$$

$$t(-a) = c, \quad t(b) = d, \quad \lambda(\varepsilon) > 0, \tag{16}$$

changes significantly when the location of one of the endpoints of the interval $[-a, b]$ suffers slight variations of order $O(\varepsilon^{\delta})$. The quantity $I = \lambda(0)(a^2 - b^2)$ characterizes the different behaviors of $t(x)$. The perturbation of the location of one endpoint causes the existence of a solution $t(x)$ with two boundary layers when $I = 0$, while for $I \neq 0$ (but $I \simeq 0$) there is a solution with only one boundary layer (the other disappears, since b considered as a function $b(a)$ is increased by a small amount).

Clearly, for $n = 0$, (15), (16) is the inverse problem corresponding to

$$\varepsilon x'' = -2\lambda(\varepsilon) x x'^2, \tag{17}$$

$$x(c) = -a, \quad x(d) = b \quad \text{where } G(b) = 0 \Leftrightarrow b = a \quad \text{with } G(b) = I. \tag{18}$$

Then, in the case $I = 0$, $t(x)$ has two boundary layers, due to the fact that the solution $x(t)$ of (17), (18) has a free layer at $b^* = a$. Actually, these two boundary layers exist for all $b \in \hat{\xi}(b^*)$, which is, in this case, the ε-gal (a).

On the contrary, if we consider $b \in$ hal (a) but outside the ε-gal (a), it turns out from Theorem 4.3.1 that $x(t)$ has a boundary layer; then $t(x)$ must have only one boundary layer. That is, the slight variations of $b = a$ that make one boundary layer disappear must be of order ε^{δ} but with $\delta < 1$. In fact, let us consider $b \simeq a$ but $b \notin \hat{\xi}(b^*)$ (for example, $b = a \pm \sqrt{\varepsilon}$). Then, if $G(b) \lesseqgtr 0$ $x(t)$ has a boundary layer at c, $t(x)$ has a boundary layer at $x = b$ while the other disappears. If $G(b) \gtreqless 0$ $x(t)$ has a boundary layer at d; thus $t(x)$ has a boundary layer only at $x = -a$.

5. Numerical results.

Example 5.1. The numerical solutions of $\varepsilon x'' = -4x x'^{3/2}$, $x(0) = -1$, $x(1) = b$, with $b \simeq 1$ and $\varepsilon = 0.05$, are shown in Fig. 5.1. Significant changes in the position of t_0 are obtained for variations of $b^* = 1$ of order 10^{-3}. The location of the transition point t_0 varies from $t_0 = 0.23$ for $b = 1.001$ to $t_0 = 0.89$ for $b = 0.9949$. The set of values of b such that there is a free layer is the $\varepsilon^{3/2}$-gal (1).

Example 5.2. This example corresponds to $\varepsilon x'' = (1 - 3x^2)x'$, $x(0) = 0$, $x(1) = b$, where $b \simeq 1$ and $\varepsilon = 0.05$. Slight variations of $b^* = 1$, less than 10^{-5}, change the position of t_0 from 0.38 to 0.73. In this case, the phenomenon is much sharper than in the first example, as it occurs for $b \in \varepsilon$-microgal (1) (see Fig. 5.2).

6. Examples.

Example 6.1. This is actually a whole family of problems

$$\varepsilon x'' = x x'^{[s]},$$

$$x(-1) = a, \qquad x(1) = b,$$

where s is a real number $s \geqq 1$ and $x'^{[s]} = x' |x'^{s-1}|$.

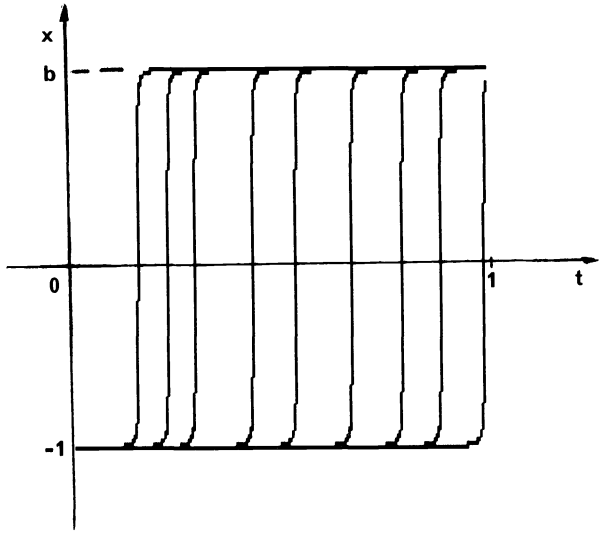


FIG. 5.1. *Solutions of* $\varepsilon x'' = -4xx'^{3/2}$, $x(0) = -1$, $x(1) = b$, *for nine values of* $b \simeq 1$ *and* $\varepsilon = 0.05$.

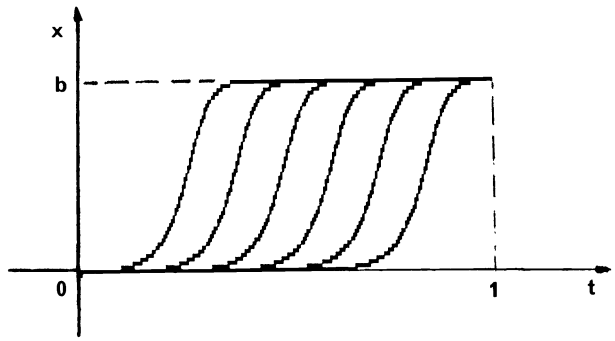


FIG. 5.2. *Solutions of* $\varepsilon x'' = (1-3x^2)x'$, $x(0) = 0$, $x(1) = b$, *for six values of* $b \simeq 1$ *and* $\varepsilon = 0.05$.

The jumps in their observability spaces are, for all $1 \leqq s \leqq 2$, infinitely close to the parabola of equation $V(x) \simeq x^2/2 + C$. Then free layers are possible only in the halo H of the straight line of equation $b = -a$.

If $1 \leqq s \leqq 2$, the theory discussed above shows that the problem has a unique solution with a free layer for b in the galaxy $\hat{\xi} \subset H$ for $a \gg 0$.

If $s > 2$, the equation is more than quadratic and, in this case, the jumps are also infinitely close to the parabolas $V(x) \simeq x^2/2 + C$, but only for $V(x)$ such that $|V(x)| \leqq V_0$, where $V_0 = \varepsilon/(s-2)$ [5]. If $s \gg 2$, well-known results [14] and a recent result of Diener [5] ensure that this problem has no solution if $a \neq b$. In fact, suppose there exists a solution—it must be a slow one, since no jump is possible as $V_0 \simeq 0$. But as slow solutions are constants, it is not possible to satisfy the boundary conditions except if $a \simeq b$. On the other hand, if $s \geqq 2$, the boundary value problem has a solution if and only if a and $b \in R$ (see Fig. 6.1). In this case, the free layers exist for values of $b \in \hat{\xi}$ for $0 \ll a \leqq \sqrt{V_0}$.

In spite of the fact that the jumps are, in the (x, V) space, the same for all $s \in [0, 2]$, our results do not apply when $0 < s < 1$ as f is not of class C^1.

The galaxy $\hat{\xi}$ and the position of t_0 depend on the values of s.

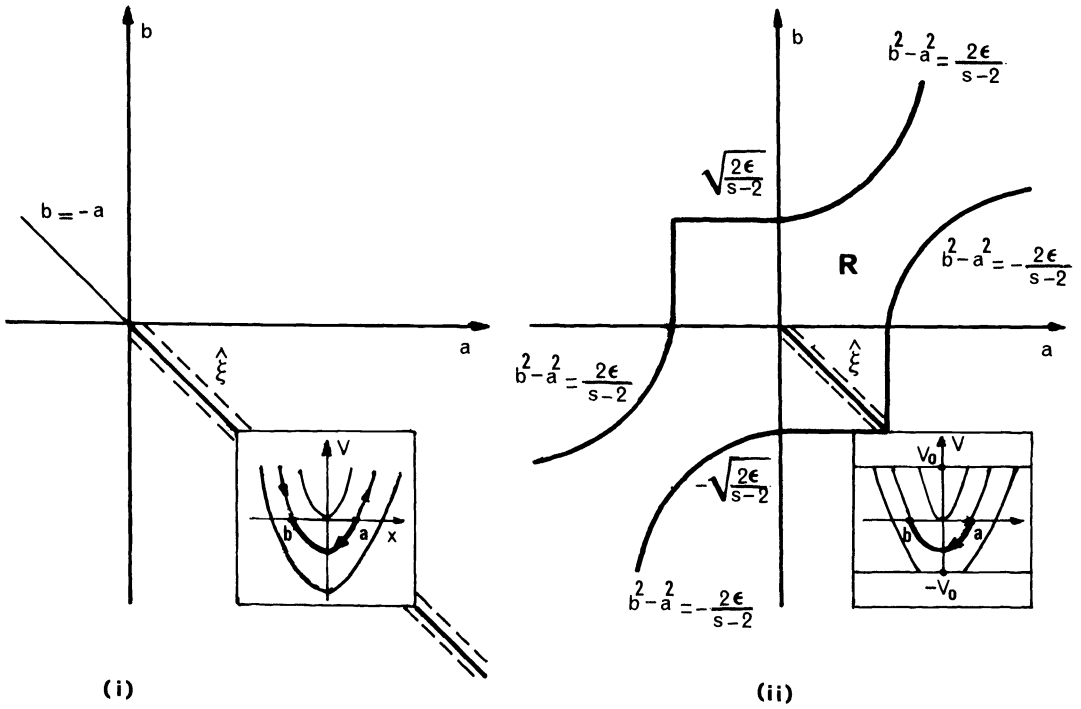


FIG. 6.1. *The $\hat{\xi}$ galaxy of the free layers of $\varepsilon x'' = xx'^{[s]}$, $x(-1) = a$, $x(1) = b$ is contained in the halo of the straight line $b = -a$ with* (i) $0 \ll a$ *if* $1 \leqq s \leqq 2$; (ii) $0 \ll a < \sqrt{V_0}$ *if* $s \geqq 2$. *In the square, the rapid portion of a trajectory associated with a solution $x(t)$ is drawn in the plane (x, V), for a value of $b \in \hat{\xi}$.*

If $s = 1$, $\hat{\xi} = \varepsilon$-microgal $(-a)$, that is, the phenomenon takes place for b exponentially close to $(-a)$ and the location is given by

$$t_0 \simeq \frac{b+a}{b-a} + \frac{\varepsilon}{b-a} \log\left(\frac{\psi(b)}{(b^2-a^2)/2\varepsilon + \psi(b)}\right),$$

with $\psi(b)$ defined by

$$\int_a^b \frac{dx}{(x^2-a^2)/2\varepsilon + \psi(b)} = d - c.$$

If $1 < s \leqq 2$ and if $s \geqq 2$, $\hat{\xi} = \varepsilon^{3-s}$-gal $(-a)$, in this case the transition occurs for b such that $b + a = O(3^{3-s})$ and

$$t_0 \simeq \frac{b+a}{b-a} + \frac{\varepsilon}{b-a}\,\frac{1}{s-1} \cdot \left(\left(\frac{1}{(2-s)(b^2-a^2)/2\varepsilon + (\psi(b))^{2-s}}\right)^{s-1/s-2} - \left(\frac{1}{\psi(b)}\right)^{s-1}\right), \qquad s \neq 2,$$

with $\psi(b)$ given by

$$\int_a^b \frac{dx}{((2-s)(b^2-a^2)/2\varepsilon + (\psi(b))^{2-s})^{1/2-s}} = d - c$$

or

$$t_0 \simeq \frac{b+a}{b-a} + \frac{\varepsilon}{b-a}\left(\frac{\exp((a^2-b^2)/2\varepsilon) - 1}{\int_a^b \exp((a^2-x^2)/2\varepsilon)\,dx}\right), \qquad s = 2.$$

Example 6.2. We now consider the problem studied by Chang and Howes [1] which arises in modelling the flow of a compressible fluid.

The problem

$$\varepsilon x'' = ((\gamma+1)/2 - 1/x^2)x', \qquad 0<t<1,$$
$$x(0) = a, \qquad x(1) = b$$

models the one-dimensional and steady-state flow that arises when a gas is injected at a supersonic velocity a in a duct of uniform cross-sectional area and a back pressure is applied. The adiabatic index γ is a constant value between 1 and $\frac{5}{3}$, x is the dimensionless velocity of the gas relative to the velocity of sound, and t is the dimensionless distance with $t=0$ at the entrance of the duct. ρ_0 is a reference density, c_0 is the velocity of sound, and $\varepsilon = \mu\gamma/\rho_0 c_0$ is infinitesimal when the coefficient of viscosity μ is infinitesimal.

Then we want to determine the subsonic velocity b at $t=1$ that produces a supersonic-subsonic shock in the duct when a supersonic velocity a is given at $t=0$ and the position of the shock.

In this case, the jump is, in its observability plane, infinitely close to

$$V(x) \simeq \left(\frac{\gamma+1}{2}x + \frac{1}{x}\right) - \left(\frac{\gamma+1}{2}a + \frac{1}{a}\right).$$

The relation $G(b^*) = \int_a^{b^*} g(x)\,dx = 0$, which is known as the Rankine-Hugoniot shock condition, is

$$\frac{\gamma+1}{2}a + \frac{1}{a} = \frac{\gamma+1}{2}b^* + \frac{1}{b^*},$$

from which it follows that b^* satisfies the well-known Prandtl relation

$$ab^* = \frac{2}{\gamma+1}.$$

Then, by virtue of Theorem 4.3.1, there is a shock in the interior of the duct, not only for $b^* = (2/\gamma+1)1/a$, but also for the values of $b \in \varepsilon$-microgal (b^*) (see Fig. 6.2).

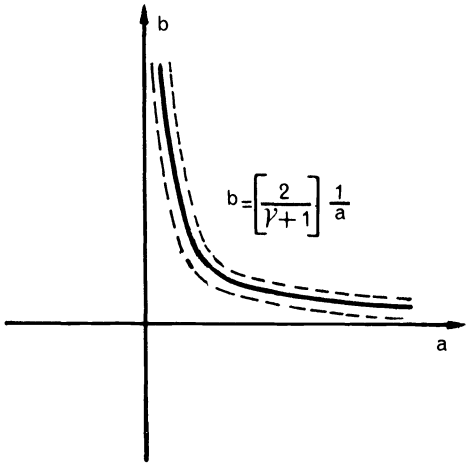


FIG. 6.2. *The $\hat{\xi}$ galaxy of the free layers of $\varepsilon x'' = (2/(\gamma+1) - 1/x^2)x'$, $x(0)=a$, $x(1)=b$ is contained in the halo of the curve of equation $b = (2/(\gamma+1)) \cdot 1/a$.*

The transition point is located at

$$t_0 \simeq \frac{a(a-b^2/b^*)}{a^2-b^2} - \frac{a^2b^2}{a^2-b^2}\,\varepsilon \log\left(\frac{\psi(b)}{G(b)/\varepsilon+\psi(b)}\right)$$

with $\psi(b)$ defined by

$$\int_a^b \frac{dx}{G(x)/\varepsilon+\psi(b)} = 1.$$

When $b = b^*$,

$$t_0 \simeq \frac{a}{a+b^*}.$$

When $b \notin \hat{\xi}$, if $b \gtrsim b^*$ the shock is always close to the end of the duct. On the contrary, if $b \lesssim b^*$ the transition point is near the entrance of the duct.

7. Appendix. In this appendix we give only the definitions of the nonstandard words and the main tools that we use in this work. For an introduction to nonstandard analysis we refer the reader to [8] and also [10], [13].

class S^0, S^1: An internal function f is S-continuous at x if for all y, $y \simeq x \Rightarrow$, $f(y) \simeq f(x)$. For example, $f(x) = \operatorname{arctg}(x)$ is S-continuous at $x = 0$ but $f(x) = \operatorname{arct}(x/\varepsilon)$, and ε infinitesimal is not S-continuous at $x = 0$. An internal function f is of class S^0 in E if $f(x)$ is near standard when $x \in E$, x is near standard, and f is S-continuous at x. A differentiable internal function f is of class S^1 if and only if f and f' are of class S^0. For example, $f(x) = 0$ if $x \leqq 0$, $f(x) = \varepsilon x$ if $x > 0$, is of class S^1 at $x = 0$, but $f(x) = \varepsilon \sin(x/\varepsilon)$ is not of class S^1 at $x = 0$ with ε infinitesimal.

Fehrele's Principle. This principle states "No halo is a galaxy." That is, if a halo contains a galaxy, that halo must overflow the galaxy.

galaxy. The set of limited reals is called the main galaxy and is denoted by $\mathbb{G}$. The pre-image of $\mathbb{G}$ by an internal function f is called a pregalaxy. It may be an internal set (e.g., if $f(x) = 1$) or an external set (e.g., if $f(x) = x/\varepsilon$, ε infinitesimal). In the last case, the pregalaxy is called a galaxy. The ε-galaxy of a, denoted by ε-gal (a), a real number, is the pre-image of $\mathbb{G}$ by $f(x) = (x-a)/\varepsilon$. ε-gal $(a) = \{x \in \mathbb{R}: x = a + \varepsilon L, L \in \mathbb{G}\}$. The ε-microgalaxy of a, denoted by ε-microgal (a), is the pre-image of $\mathbb{G}^+$ by $f(x) = -1/(\varepsilon \log|x-a|)$. ε-microgal $(a) = \{x \in \mathbb{R}: |x-a| < \exp(-1/L^2\varepsilon),\ L \in \mathbb{G}\}$.

halo. The set of infinitesimals is called the halo of zero and it is denoted by hal (0). The pre-image of hal (0) by an internal function f is called a prehalo. It may be an internal set (if, e.g., f is constant) or an external set (if, e.g., $f(x) = x/\varepsilon$) and, in this case, the prehalo is called a halo. The ε-halo of a, denoted by ε-hal (a), is the pre-image of hal (0) by $f(x) = (x-a)/\varepsilon$. The ε-microhalo of a, denoted by ε-microhal (a), is the pre-image of the halo (0) by $f(x) = 1/(\varepsilon \log|x-a|)$. The complement of a halo in an internal set is a galaxy.

infinitely close. The real numbers a and b are infinitely close if $a - b = \eta$, η infinitesimal. We set them $a \simeq b$ if a and b are infinitely close and $a \not\simeq b$ if not. The functions f and g are infinitely close in E if for any $x \in E$ $f(x) \simeq g(x)$. We denote by $a \ll b$ $(a \gg b)$ if $a < b$ $(a > b)$ and $a \not\simeq b$.

infinitesimal. A real number ε is infinitesimal if its absolute value is smaller than any positive standard real number.

internal–external. We work with two kinds of sets: the usual sets of the Zermelo-Fraenkel theory, which are called internal sets (for example, $\mathbb{N}$, $\mathbb{R}$, $(0, w)$) and the sets built up with the help of the new predicate "standard" (or with one of those derived from it, such as "infinitesimal," "large," or "limited"). These sets are called external or strictly external when at least one classical theorem does not apply to them. For example, the sets of infinitesimal or of standard reals are external (they are bounded

subsets of $\mathbb{R}$ with no upper bound). A function f is called internal if its graph is an internal set. For example, $f(x) = \varepsilon x$, ε infinitesimal, is internal; however, $f(x) = 0$ if x is infinitesimal and $f(x) = 1$ if not is an external function.

limited. A real number L is limited if its absolute value is smaller than some standard integer. Any limited real number is infinitely close to a unique real standard, called its standard part and denoted by $^\circ L$.

near. The use of this adverb in an expression such as "near equation" or "near standard" means that it differs from being an equation or a standard by an infinitesimal.

near standard. A real number x is near standard if there exists a standard s such that $x \simeq s$.

S-continuous. See "class of $S^0 - S^1$."

standard. The adjective standard is a new predicate introduced in the mathematical language. The use of it is governed by three axioms: transfer, idealization, and standardization principles (see [8], [10]). The sets may be standard or not. The sets $\mathbb{R}$, $\mathbb{Q}$, $[0, 1]$, $\Phi(=0)$, $\{\Phi\} = 1$, arct (x) are standard; actg x/ε and $[0, 1/\varepsilon]$ are nonstandard if ε is infinitesimal, $\varepsilon \neq 0$.

standard part. The standard part of a real x, denoted by $^\circ x$, is the unique standard (when it exists) that is infinitely close to x. The standard part of a function f is the unique standard function $(^\circ f)$, when it exists, such that $f(x) \simeq {}^\circ f(x)$ for any x near standard in the domain of f.

unlimited. A real number W is unlimited if it is larger than any standard integer.

REFERENCES

[1] K. CHANG AND F. HOWES, *Nonlinear Singular Perturbation Phenomena: Theory and Application*, Springer-Verlag, Berlin, New York, 1984.

[2] L. CROCCO, *A suggestion for the numerical solution of the steady Navier-Stokes equations*, AIAA J., 3 (1965), pp. 1824-1832.

[3] F. DIENER, *Méthode du plan d'observabilité*, Thèse, Université Louis Pasteur, Strasbourg, France, 1981.

[4] ———, *Sauts des solutions des équations $\varepsilon x'' = f(t, x, x')$*, SIAM J. Math. Anal., 17 (1986), pp. 533-559.

[5] ———, *Equations surquadratiques et disparition des sauts*, SIAM J. Math. Anal. 19 (1988), pp. 1127-1134.

[6] F. HOWES, *Boundary-interior layer interactions in nonlinear singular perturbation theory*, Mem. Amer. Math. Soc., 203 (1978), pp. 1-108.

[7] L. K. JACKSON, *Subfunctions and second order ordinary differential inequalities*, Adv. in Math., 2 (1968), pp. 307-363.

[8] R. LUTZ AND M. GOZE, *Non-Standard Analysis. A Practical Guide With Applications*, Lecture Notes in Math. 881, Springer-Verlag, Berlin, New York, 1981.

[9] B. J. MATKOWSKY, *On boundary layer problems exhibiting resonance*, SIAM Rev., 17 (1975), pp. 82-100.

[10] E. NELSON, *Internal set theory*, Bull. Amer. Math. Soc., 83 (1977), pp. 1165-1198.

[11] R. E. O'MALLEY, *Introduction to Singular Perturbations*, Academic Press, New York, 1974.

[12] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[13] A. ROBINSON, *Non-Standard Analysis*, North-Holland, Amsterdam, New York, 1966.

[14] M. VISIK AND L. LIUSTERNIK, *Initial jump for nonlinear differential equations containing a small parameter*, Soviet Math. Dokl., 1 (1960), pp. 749-752.

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1281-1304, September 1990
© 1990 Society for Industrial and Applied Mathematics
011

NONCONVEX FUNCTIONALS RELATED TO MULTIPHASE SYSTEMS*

AUGUSTO VISINTIN†

Abstract. Let Ω be a bounded domain of $\mathbb{R}^N (N \geqq 1)$, with ϕ a (nonconvex) lower semicontinuous function $\mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, such that for any $u \in L^1(\Omega)$, $\Phi(u) := \int_\Omega \phi(u(x))\, dx > -\infty$.

Let $\Lambda : L^1(\Omega) \to [0, +\infty]$ fulfill the *generalized co-area formula* $\Lambda(u) = \int_{\mathbb{R}} \Lambda(H(u-s))\, ds (\leqq +\infty)$ for all $u \in L^1(\Omega)$, where $H(\xi) = 0$ if $\xi < 0$, $H(\xi) = 1$ if $\xi \geqq 0$. For instance,

$$V(u) := \int_\Omega |\nabla u| = \sup \left\{ \int_\Omega u \operatorname{div} \eta \, dx : \eta \in C_c^1(\Omega)^N, |\eta| \leqq 1 \right\},$$

$$\Lambda_r(u) := \int\int_{\Omega^2} |u(x) - u(y)|\, |x-y|^{-(N+r)}\, dx\, dy \qquad (0 < r < 1),$$

$$\tilde{\Lambda}_r(u) := \int_{\mathbb{R}^+} h^{-(1+r)}\, dh \int_\Omega (\operatorname{ess\,sup}_{B_h(x) \cap \Omega} u - \operatorname{ess\,inf}_{B_h(x) \cap \Omega} u)\, dx \qquad (0 < r < 1),$$

where $B_h(x) := \{y : |y - x| \leqq h\}$. Here it is proven that if $\Lambda = \Lambda^{**}$, then for any $u \in L^1(\Omega)$, $\partial(\Phi + \Lambda)(u) = \partial\Phi(u) + \partial\Lambda(u)$ in $L^\infty(\Omega)$, and $(\Phi + \Lambda)^{**}(u) = \Phi^{**}(u) + \Lambda(u)$.

This and another result entail that for any $\xi \in L^\infty(\Omega)$, if u is an absolute (relative, respectively) minimum of $\Psi_\xi : v \mapsto \int_\Omega [\phi(v(x)) - \xi(x)v(x)]\, dx + \Lambda(v)$ in $L^1(\Omega)$, then there exists $\tilde{\xi} \in L^\infty(\Omega)$ such that almost everywhere in Ω, $u(x)$ is an absolute (relative, respectively) minimum of $y \mapsto \phi(y) - \tilde{\xi}(x)y$ in $\mathbb{R}$. Hence, for both sorts of minima, certain values are a priori excluded from $u(\Omega)$, which can be *nonconvex*. This can represent the occurrence of a *phase structure*, i.e., *pattern formation*. If ϕ is the free-energy density function of some substance, Λ can model the phase interaction contribution to the global free energy. The absolute and relative minima of Ψ_ξ are related to the *stable* and *metastable equilibrium* states, respectively.

Solid-liquid systems are discussed in particular. The proposed model accounts for *supercooling* and *superheating* effects. If $\Lambda = V$, the mean curvature of the solid-liquid interface $\mathcal{S}$ is as prescribed by the *Gibbs-Thomson law*. If $\Lambda = \Lambda_r$ $(0 < r < 1)$, $\mathcal{S}$ can be more irregular, as in dendritic formations and snowflakes. This model can be extended to include *mushy regions*.

Key words. surface tension, co-area formula, absolute and relative minima, stable and metastable states

AMS(MOS) subject classifications. 49A29, 80A00

1. Introduction and presentation of the model.

This paper consists of two parts. In §§ 2 and 3 a certain class of nonconvex functionals is studied; in §§ 4 and 5 these results are applied to model stationary multiphase systems. In this section we will outline these developments moving from the physical aspects.

Substances capable of attaining two phases are characterized by a nonconvex dependence of the free-energy density on the state variables. Here we consider the potential function ϕ sketched in Fig. 1, although the developments of §§ 3 and 4 hold for a general (nonconvex) lower semicontinuous function $\mathbb{R} \to \mathbb{R} \cup \{+\infty\}$. To explore all the possible stationary configurations, we will use a linearly perturbed potential, so for any $\xi, u \in \mathbb{R}$ we set $\phi_\xi(u) := \phi(u) - \xi u$. This construction has a physical meaning—for instance, in solid-liquid systems ξ is proportional to the relative temperature. Let our system occupy a bounded domain $\Omega \subset \mathbb{R}^N$ $(N \geqq 1)$; for any $\xi \in L^\infty(\Omega)$ and any $u \in L^1(\Omega)$, we set $\Phi(u) := \int_\Omega \phi(u(x))\, dx$ $(\leqq +\infty)$ and $\Phi_\xi(u) := \int_\Omega \phi_{\xi(x)}(u(x))\, dx$. Stationary configurations of a thermodynamic system have either a *stable* or a *metastable*

* Received by the editors December 21, 1988; accepted for publication November 6, 1989. The results of the present paper were announced at the colloquium on free boundary problems held in Irsee in June 1987 [22].

† Dipartimento di Matematica dell'Università degli Studi di Trento, 38050 Povo (Trento), Italy, and Istituto di Analisi Numerica del C.N.R., C.so C. Alberto 5, 27100 Pavia, Italy.

equilibrium. Stable states can be attained for arbitrarily long times, whereas metastable ones will eventually decay because of fluctuations. An example of the latter is given by a supercooled liquid. If the system is governed by a potential, the stable states correspond to its absolute minima, and the metastable ones to its relative (nonabsolute) minima.

For any $\xi \in L^\infty(\Omega)$, obviously $u \in L^1(\Omega)$ is an absolute minimum of the functional Φ_ξ in $L^1(\Omega)$ if and only if almost everywhere in Ω, $u(x)$ is an absolute minimum of the function $\phi_{\xi(x)}$ in $\mathbb{R}$. But Φ_ξ has no relative (nonabsolute) minimum in $L^1(\Omega)$ for any $\xi \in L^\infty(\Omega)$, even if almost everywhere in Ω, $\phi_{\xi(x)}$ has a relative minimum in $\mathbb{R}$ (Proposition 5 in § 4). This excludes the possibility of representing any metastable state by means of the potential Φ_ξ.

We remedy this drawback by including a *space interaction* term in the potential functional. We exclude any term of the form $\Lambda^p(u) := \int_\Omega |\nabla u|^p \, dx$ $(\leqq +\infty)$ with $l \leqq p \leqq \infty$, because, by well-known trace theorems, the condition $\Lambda^p(u) < +\infty$ is not consistent with the discontinuities along surfaces that occur in multiphase states. Rather, we will consider interaction functionals of the form

$$V(u) := \int_\Omega |\nabla u| = \sup \left\{ \int_\Omega u \operatorname{div} \eta \, dx : \eta \in C_c^1(\Omega)^N, |\eta| \leqq 1 \right\}, \tag{1.1}$$

$$\Lambda_r(u) := \int\int_\Omega |u(x) - u(y)| \, |x-y|^{-(N+r)} \, dx \, dy \qquad (0 < r < 1), \tag{1.2}$$

$$\tilde{\Lambda}_r(u) := \int_{\mathbb{R}^+} h^{-(1+r)} \, dh \int_\Omega (\operatorname*{ess\,sup}_{B_h(x)\cap\Omega} u - \operatorname*{ess\,inf}_{B_h(x)\cap\Omega} u) \, dx \qquad (0 < r < 1), \tag{1.3}$$

where $B_h(x) := \{y : |y-x| \leqq h\}$. Thus Dom $(V_1) = BV(\Omega)$: Banach space of functions $\Omega \to \mathbb{R}$ with bounded total variation; Dom $(\Lambda_r) = W^{r,1}(\Omega)$: fractional Sobolev space [1].

Let us set $H(\xi) = 0$ if $\xi < 0$, $H(\xi) = 1$ if $\xi \geqq 0$. Any of these functionals fulfills the *generalized co-area formula*

$$\Lambda(u) := \int_{\mathbb{R}} \Lambda(H(u-s)) \, ds (\leqq +\infty) \quad \forall u \in L^1(\Omega). \tag{1.4}$$

The implications of this property are studied in a more general context in § 2; there $GC(\Omega)$ is defined to be the set of functionals $\Lambda : L^1(\Omega) \to [0, +\infty]$ that fulfill (1.4). This class of functionals is studied in more detail in [21].

For any $\Lambda \in GC(\Omega)$ and any $\xi \in L^\infty(\Omega)$, we then introduce the potential functional $\Psi_\xi := \Phi_\xi + \Lambda$. Of course Ψ_ξ is compatible with a multiphase structure, in which discontinuities occur on surfaces separating two different phases. Moreover, for $\Lambda = V$ or $\Lambda = \Lambda_r$ or $\Lambda = \tilde{\Lambda}_r$ $(0 < r < 1)$, Ψ_ξ can also have relative (nonabsolute) minima in $L^1(\Omega)$, for suitable φ and ξ.

In Theorem 3 of § 3 we show that for any lower semicontinuous and convex $\Lambda \in GC(\Omega)$

$$(\Phi + \Lambda)^{**} = \Phi^{**} + \Lambda \quad \text{in } L^1(\Omega), \tag{1.5}$$

$$\partial(\Phi + \Lambda) = \partial\Phi + \partial\Lambda \quad \text{in } L^1(\Omega) \tag{1.6}$$

(here F^{**} denotes the lower semicontinuous convex hull of F [4]).

Equation (1.6) entails that, for any $\xi \in L^\infty(\Omega)$, if u is an *absolute* minimum of Ψ_ξ in $L^1(\Omega)$, then $\partial\phi(u(x)) \neq \varnothing$ almost everywhere in Ω; that is, for instance, for ϕ as in Fig. 1, u has the *phase structure* $u(\Omega) \subset \mathbb{R} \setminus]b, g[$—note that this is a *nonconvex constraint.* This can represent several phenomena of *pattern formation.*

Theorem 4 in § 3 proves a similar result for *relative* minima: Still, for any lower semicontinuous and convex $\Lambda \in GC(\Omega)$ and any $\xi \in L^\infty(\Omega)$, if u is either a relative or

an absolute minimum of $\Psi_\xi := \Phi_\xi + \Lambda$ in $L^1(\Omega)$, then "$\partial_{\text{loc}}\phi(u(x)) \neq \varnothing$" almost everywhere in Ω; the *local subdifferential* ∂_{loc} is defined in (3.48). For ϕ as in Fig. 1, such a u has the phase structure $u(\Omega) \subset \mathbb{R}\backslash[c, f]$; note also that this constraint is *nonconvex*, and can represent *pattern formation* phenomena.

For a solid-liquid system we can assume ϕ as sketched in Fig. 2, and ξ proportional to the relative temperature; here $u = -1$ corresponds to the solid and $u = 1$ to the liquid. Note that $u = -1$ is an absolute (relative, respectively) minimum of ϕ_ξ for $\xi \leqq 0$ ($0 < \xi < \phi'(-1)$, respectively); and that $u = 1$ is an absolute (relative, respectively) minimum of ϕ_ξ for $\xi \geqq 0$ ($\phi'(1) < \xi < 0$, respectively).

For a space distributed system we consider a potential functional of the form $\Psi_\xi := \Phi_\xi + \Lambda$, with Λ proportional either to V or to Λ_r ($0 < r < 1$); here we neglect any exterior boundary contribution. Then the absolute minima of Ψ_ξ can be interpreted as the stable states, and the relative minima as the metastable ones. This gives a representation of *supercooled* and *superheated* states, corresponding to $\xi < 0$ in the liquid and to $\xi > 0$ in the solid (respectively), and of the thresholds for *nucleation* of a new phase.

If we choose $\Lambda = \sigma V$, σ denoting the surface tension coefficient, then the extremal points of Ψ_ξ contain a weak form of the classical *Gibbs-Thomson law* at the solid-liquid interface $\mathcal{S}$:

$$\xi = -2\sigma\kappa \quad \text{on } \mathcal{S} \tag{1.7}$$

for $\Omega \subset R^3$; here ξ is assumed continuous (at least on $\mathcal{S}$) and κ denotes the local mean curvature of $\mathcal{S}$, assumed positive for a solid ball [20]. If we take $\Lambda = \text{constant}$; Λ_r, $0 < r < 1$, then $\mathcal{S}$ can have a *dendritic shape*, similar to that of snowflakes; and the smaller r is, the more irregular $\mathcal{S}$ can be.

Suitably modified profiles of φ (cf. Fig. 3) allow us to represent intermediate phases as well. These include *mushy regions* and, for liquid-vapor systems, *clouds* and *fog*. Finally, the phase interaction functionals V and Λ_r ($0 < r < 1$) are interpreted through a *quasi-chemical* approach to multiphase systems (cf. [3, Chap. 2] and [17]).

Bibliographical note. Several mathematical papers on surface-tension effects in two-phase systems have appeared in recent years. Some authors have used the van der Waals/Cahn-Hilliard *phase-field model*, which is characterized by a continuous potential function $\phi : \mathbb{R} \to \mathbb{R}$ and by an interaction term proportional to $\int_\Omega |\nabla u|^2\, dx$ [2].

Other models based on the functional V (cf. (1.1)) have been studied by Gurtin [10]-[12] and Visintin [18], [20]. In particular, in [20] the potential function ϕ sketched in Fig. 2 is proposed for representing solid-liquid systems. Modica [14], [15] has shown how the model based on V can be retrieved from a phase-field model by means of a Γ-limit in the sense of De Giorgi.

The evolution case, namely, the formation and movement of interfaces (*pattern evolution*), is studied in [23] using a *hysteresis* model; here we note only that the presence of metastable states in the stationary case corresponds to hysteresis effects in the evolution.

2. Generalized co-area formula. Let Ω be a set of $\mathbb{R}^N$ ($N \geqq 1$). We set

$$H_s(y) = \begin{cases} 0 & \text{if } y < s, \\ 1 & \text{if } y \geqq s. \end{cases}$$

DEFINITION 1. We will denote by $GC(\Omega)$ the class of functionals $\Lambda : L^1(\Omega) \to [0, +\infty]$ that are proper (i.e., $\Lambda \not\equiv +\infty$), and that fulfill the *generalized co-area formula*

$$\Lambda(u) = \int_{\mathbb{R}} \Lambda(H_s(u))\, ds \ (\leqq +\infty) \quad \forall u \in L^1(\Omega), \tag{2.1}$$

with the convention that the integral is set equal to $+\infty$ if the function $\mathbb{R}\to[0,+\infty]: s\mapsto\Lambda(H_s(u))$ is not measurable.

For any $\Lambda\in GC(\Omega)$, we also set $\mathrm{Dom}\,(\Lambda):=\{u\in L^1(\Omega):\Lambda(u)\neq+\infty\}$ and $\hat\Lambda(u):=\Lambda(-u)$ for all $u\in L^1(\Omega)$; note that $\hat\Lambda\in GC(\Omega)$.

PROPOSITION 1. *For any* $\Lambda\in GC(\Omega)$,

$$\Lambda(u+c)=\Lambda(u)\quad\forall u\in L^1(\Omega),\quad\forall c\in\mathbb{R}, \tag{2.2}$$

$$\Lambda(cu)=c\Lambda(u)\quad\forall u\in L^1(\Omega),\quad\forall c>0, \tag{2.3}$$

$$\Lambda(c)=0\quad\forall c\in\mathbb{R}, \tag{2.4}$$

$$\Lambda(u)=\Lambda(u\wedge c)+\Lambda(u\vee c)\quad\forall u\in L^1(\Omega),\quad\forall c\in\mathbb{R} \tag{2.5}$$

(*where* $(u\wedge c)(x):=\min(u(x),c)$, $(u\vee c)(x):=\max(u(x),c)$),

(2.6) *If* Λ *is convex, then* $\mathrm{Dom}\,(\Lambda)$ *is a convex cone,* $\mathrm{Dom}\,(\Lambda+\hat\Lambda)$ *is a linear subspace of* $L^1(\Omega)$ *and* $\Lambda+\hat\Lambda$ *is a seminorm.*

The proof of these properties is straightforward. □

Examples of functionals of $GC(\Omega)$:

(i) Trivial cases: $\Lambda^0\equiv 0$; $\Lambda_c(u)=0$ if $u=$ constant almost everywhere in Ω; $\Lambda_c(u)=+\infty$ otherwise.

(ii) We set

$$V(u):=\int_\Omega|\nabla u|:=\sup\left\{\int_\Omega u\,\mathrm{div}\,\eta\,dx:\eta\in C_c^1(\Omega)^N,|\eta|\leqq 1\right\}(\leqq+\infty)\quad\forall u\in L^1(\Omega); \tag{2.7}$$

then (2.1) holds and coincides with the classical Fleming-Rishel co-area formula [6], [8, p. 20]. $\mathrm{Dom}\,(V)=BV(\Omega)$, a Banach space of integrable functions with bounded total variation. V is convex and lower semicontinuous in $L^1(\Omega)$; that is, $V=V^{**}$.

(iii) The functional

$$\Lambda_{\mathrm{osc}}(u):=\operatorname*{ess\,osc}_\Omega u\ (:=\operatorname*{ess\,sup}_\Omega u-\operatorname*{ess\,inf}_\Omega u)\quad(\leqq+\infty)\quad\forall u\in L^1(\Omega). \tag{2.8}$$

(iv) For any measurable function $g:\Omega^2\to\mathbb{R}^+$, we set

$$\Lambda_g(u):=\iint_{\Omega^2}|u(x)-u(y)|g(x,y)\,dx\,dy\quad(\leqq+\infty)\quad\forall u\in L^1(\Omega). \tag{2.9}$$

It is easy to check that $\Lambda_g\in GC(\Omega)$, by the identity

$$|\xi-\eta|=\int_{\mathbb{R}}|H_s(\xi)-H_s(\eta)|\,ds\quad\forall\xi,\eta\in\mathbb{R}$$

(here applied with $\xi=u(x)$, $\eta=u(y)$), and by Fubini's theorem. Λ_g is also convex and, by Fatou's lemma, is lower semicontinuous in $L^1(\Omega)$; that is, $\Lambda_g=\Lambda_g^{**}$.

(v) As a particular case of example (iv), we take

$$g_r(x,y):=|x-y|^{-(N+r)}\quad\forall x,y\in\Omega\ (x\neq y),\quad\forall r\in\,]0,1[\tag{2.10}$$

and set $\Lambda_r:=\Lambda_{g_r}$. This is the standard seminorm of the fractional Sobolev space $W^{r,1}(\Omega)$ $(=\mathrm{Dom}\,(\Lambda_r))$.

(vi) For any measurable function $f:\Omega\times\mathbb{R}^+\to\mathbb{R}^+$, we set

$$\tilde\Lambda_f(u):=\iint_{\Omega\times\mathbb{R}^+}\left(\operatorname*{ess\,osc}_{\Omega\cap B_h(x)}u\right)\cdot f(x,h)\,dx\,dh\ (\leqq+\infty)\quad\forall u\in L^1(\Omega), \tag{2.11}$$

where $B_h(x) := \{y \in \mathbb{R}^N : |x-y| < h\}$. It is easy to check that $\tilde{\Lambda}_f \in GC(\Omega)$, by example (iii) and by Fubini's theorem. $\tilde{\Lambda}_f$ is convex and by Fatou's lemma, also lower semicontinuous in $L^1(\Omega)$, as is Λ_{osc}; that is, $\tilde{\Lambda}_f = \tilde{\Lambda}_f^{**}$.

(vii) As a particular case of example (vi), we take

$$f_r(x, h) := h^{-(1+r)} \quad \text{a.e. for } (x, h) \in \Omega \times \mathbb{R}, \quad \forall r \in]0, 1[, \tag{2.12}$$

and set $\tilde{\Lambda}_r := \tilde{\Lambda}_{f_r}$. Then

$$\Lambda_r(u) \leqq \text{constant} \cdot \tilde{\Lambda}_r(u) \quad \forall u \in L^1(\Omega), \quad \forall r \in]0, 1], \tag{2.13}$$

whence $\text{Dom}(\tilde{\Lambda}_r) \subset \text{Dom}(\Lambda_r)$.

PROPOSITION 2 [21]. *If either $\Lambda = \Lambda_{osc}$, or $\Lambda = \Lambda_g$, or $\Lambda = \tilde{\Lambda}_f$, then $\Lambda \in GC(\Omega)$ and $\Lambda = \Lambda^{**}$. Moreover, if either $\Lambda = V$, or $\Lambda = \Lambda_r$, or $\Lambda = \tilde{\Lambda}_r$ $(0 < r < 1)$, then the injection of the Banach space* Dom (Λ) *into $L^1(\Omega)$ is compact, provided that Ω fulfills the regularity assumptions of the classical Rellich compactness theorem.*

3. On a class of nonconvex functionals. Let Ω be a bounded domain of $\mathbb{R}^N$ $(N \geqq 1)$ endowed with the ordinary Lebesgue measure μ. Let us fix a function

$$\phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\} \quad \text{proper and lower semicontinuous;} \tag{3.1}$$

note that we do not require ϕ to be convex. We also assume that

$$\exists C_1, C_2 \in \mathbb{R}^+ : \quad \forall y \in \mathbb{R}, \qquad \phi(y) \geqq -C_1|y| - C_2, \tag{3.2}$$

whence

$$-\infty < \Phi(v) := \int_\Omega \phi(v)\, dx \quad (\leqq +\infty) \quad \forall v \in L^1(\Omega). \tag{3.3}$$

We fix a functional $\Lambda : L^1(\Omega) \to [0, +\infty]$ and any $\xi \in L^\infty(\Omega)$; we then set

$$\begin{aligned} &\phi_{\xi(x)}(y) := \phi(y) - \xi(x)y \quad \forall y \in \mathbb{R} \quad \text{a.e. in } \Omega, \\ &\Phi_\xi(v) := \int_\Omega \phi_{\xi(x)}(v(x))\, dx \quad \forall v \in L^1(\Omega), \\ &\Psi := \Phi + \Lambda, \quad \Psi_\xi := \Phi_\xi + \Lambda \quad \text{in } L^1(\Omega). \end{aligned} \tag{3.4}$$

LEMMA 1. (i) *For any $p \in [1, +\infty]$, any $f \in L^p(\Omega)$ with $f \geqq a$ $(a \in \mathbb{R})$, and any $g \in L^{p'}(\Omega)$ $(1/p + 1/p' = 1)$,*

$$\int_\Omega fg\, dx = \int_a^{+\infty} ds \int_\Omega H_s(f(x))g(x)\, dx + a \int_\Omega g\, dx. \tag{3.5}$$

Proof. Let H_s be defined as in § 2. By the identity

$$y = \int_0^{+\infty} H_s(y)\, ds \quad \forall y \in \mathbb{R}^+,$$

and by Fubini's theorem, for any measurable functions $\tilde{f}$, $g : \Omega \to \mathbb{R}^+$ we have

$$\int_\Omega \tilde{f}g\, dx = \int_0^{+\infty} ds \int_\Omega H_s(\tilde{f}(x))g(x)\, dx \quad (\leqq +\infty).$$

So, taking $\tilde{f} = f - a$, we get (3.5) for any $f \in L^p(\Omega)$, $f \geqq a$, and any $g \in L^{p'}(\Omega)$, $g \geqq 0$. It easily follows that the restriction on the sign of g can be eliminated. □

LEMMA 2. *Let $\Lambda \in GC(\Omega)$, $\{A_j\}_{j\in I}$ be a partition of $\mathbb{R}$ into intervals, possibly including one or two half-lines, and $J \subset I$. Let $u, v \in \mathrm{Dom}(\Lambda)$ be such that*

$$u = v \quad \text{a.e. in } u^{-1}(A_j) \quad \forall j \in I\setminus J, \tag{3.6}$$

$$u^{-1}(A_j) = v^{-1}(A_j) \quad \forall j \in J. \tag{3.7}$$

Then

$$\Lambda(u) - \Lambda(v) = \sum_{j\in J} \int_{A_j} [\Lambda(H_s(u)) - \Lambda(H_s(v))]\, ds. \tag{3.8}$$

Moreover, for any $\lambda \in L^\infty(\Omega)$

$$\int_\Omega \lambda(u-v)\, dx = \sum_{j\in J} \int_{A_j} ds \int_{u^{-1}(A_j)} \lambda(x)[H_s(u(x)) - H_s(v(x))]\, dx. \tag{3.9}$$

Proof. (i) Equations (3.6) and (3.7) entail

$$H_s(u) = H_s(v) \quad \text{a.e. in } \Omega \quad \forall s \in \bigcup_{j\in I\setminus J} A_j;$$

hence from (2.1)

$$\begin{aligned}\Lambda(u) - \Lambda(v) &= \sum_{j\in I} \int_{A_j} [\Lambda(H_s(u)) - \Lambda(H_s(v))]\, ds \\ &= \sum_{j\in J} \int_{A_j} [\Lambda(H_s(u)) - \Lambda(H_s(v))]\, ds.\end{aligned}$$

By the identity

$$\xi_1 - \xi_2 = \int_{\mathbb{R}} [H_s(\xi_1) - H_s(\xi_2)]\, ds \quad \forall \xi_1, \xi_2 \in \mathbb{R},$$

and by Fubini's theorem we have

$$\begin{aligned}\int_\Omega \lambda(u-v)\, dx &= \sum_{j\in I} \int_{u^{-1}(A_j)} \lambda(u-v)\, dx = \sum_{j\in J} \int_{u^{-1}(A_j)} \lambda(u-v)\, dx \\ &= \sum_{j\in J} \int_{u^{-1}(A_j)} dx\, \lambda(x) \int_{\mathbb{R}} [H_s(u(x)) - H_s(v(x))]\, ds \\ &= \sum_{j\in J} \int_{A_j} ds \int_{u^{-1}(A_j)} \lambda(x)[H_s(u(x)) - H_s(v(x))]\, dx. \qquad \square\end{aligned}$$

THEOREM 1. *Let $\Lambda \in GC(\Omega)$ (cf. § 2) and ϕ fulfill* (3.1) *and* (3.2); *set* (3.3) *and* (3.4). *Let $\{]a_i, b_i[\subset \mathbb{R}\}_{i\in I}$ be any collection of disjoint bounded open intervals, such that*

$$\phi(y) \geqq l_i(y) := \frac{\phi(b_i) - \phi(a_i)}{b_i - a_i}(y - a_i) + \phi(a_i) \quad \forall y \in]a_i, b_i[, \quad \forall i \in I. \tag{3.10}$$

Then for any $u \in L^1(\Omega)$ such that $\partial\Psi(u) \neq \varnothing$ in $L^\infty(\Omega)$

(i) *There exists a $\hat{u} \in L^1(\Omega)$ such that, setting $A := \bigcup_{i\in I}]a_i, b_i[$,*

$$\text{If } u(x) \notin A \text{ then } \hat{u}(x) = u(x) \quad \text{a.e. in } \Omega, \tag{3.11}$$

$$\hat{u}(x) \notin A \quad \text{a.e. in } \Omega, \tag{3.12}$$

$$\partial\Psi(u) \subset \partial\Psi(\hat{u}) \quad \text{in } L^\infty(\Omega). \tag{3.13}$$

(ii) *Moreover,*

$$\mu(\{x\in\Omega: a_i<u(x)<b_i,\ \phi(u(x))>l_i(u(x))\})=0 \quad \forall i\in I. \tag{3.14}$$

Remark. Part (i) can be restated as follows. For any $\xi\in L^\infty(\Omega)$, if $u\in L^1(\Omega)$ is an absolute minimum of $v\mapsto\Psi(v)-\int_\Omega \xi v\,dx$ (i.e., $\xi\in\partial(\Phi+\Lambda)(u)$), then there exists another absolute minimum $\hat u\in L^1(\Omega)$ (i.e., $\xi\in\partial(\Phi+\Lambda)(\hat u)$), which also fulfills (3.11) and (3.12).

Proof. (i) Let us fix any $\xi\in\partial\Psi(u)$. First we consider the case of A reduced to a single interval $]a, b[$. We set

$$\lambda_1 := [\phi(b)-\phi(a)]/(b-a), \qquad \lambda_2 := \phi(a),$$

$$\tilde\phi(y) := \begin{cases} \phi(y) & \forall y\in\mathbb{R}\setminus]a, b[,\\ \lambda_1(y-a)+\lambda_2 & \forall y\in]a, b[,\end{cases}$$

and define $\tilde\Phi$, $\tilde\Psi$, $\tilde\Psi_\xi$ as in (3.3) and (3.4), with ϕ replaced by $\tilde\phi$. We also set $Z := \{x\in\Omega: a<u(x)<b\}$ and

$$B_v(s) := \int_Z [\lambda_1-\xi(x)]H_s(v(x))\,dx+\Lambda(H_s(v)) \quad \forall v\in L^1(\Omega), \quad \forall s\in\mathbb{R}.$$

For any $v\in L^1(\Omega)$ such that $a\leqq v\leqq b$ almost everywhere in Z, by (2.1) and Lemma 1 we have

$$\begin{aligned}
\tilde\Psi_\xi(v) &= \int_{\Omega\setminus Z}\phi_{\xi(x)}(v(x))\,dx+\int_Z\{[\lambda_1-\xi(x)]v(x)-\partial\lambda_1+\lambda_2\}\,dx+\Lambda(v)\\
&= \int_{\Omega\setminus Z}\phi_{\xi(x)}(v(x))\,dx+\int_a^b ds\int_Z[\lambda_1-\xi(x)]H_s(v(x))\,dx\\
&\quad -a\int_Z\xi(x)\,dx+\lambda_2\mu(Z)+\left(\int_{-\infty}^a ds+\int_a^b ds+\int_b^{+\infty} ds\right)\Lambda(H_s(v))\\
&= \int_{\Omega\setminus Z}\phi_{\xi(x)}(v(x))\,dx+\int_a^b B_v(s)\,ds-a\int_Z\xi(x)\,dx\\
&\quad +\lambda_2\mu(Z)+\left(\int_{-\infty}^a ds+\int_b^{+\infty} ds\right)\Lambda(H_s(v)).
\end{aligned} \tag{3.15}$$

We note that there exists an $\hat s\in]a, b[$ such that

$$B_u(\hat s)\leqq\frac{1}{b-a}\int_a^b B_u(s)\,ds; \tag{3.16}$$

then we set

$$\hat u(x) := \begin{cases} u(x) & \text{if } x\in\Omega\setminus Z,\\ a & \text{if } x\in Z \text{ and } a\leqq u(x)<\hat s,\\ b & \text{if } x\in Z \text{ and } \hat s\leqq u(x)\leqq b.\end{cases} \tag{3.17}$$

Thus (3.11) and (3.12) hold (for $A=]a, b[$); hence

$$\begin{aligned}
&\phi_\xi(\hat u)=\tilde\phi_\xi(\hat u) \quad \text{a.e. in } \Omega,\\
&\Psi_\xi(\hat u)=\tilde\Psi_\xi(\hat u).
\end{aligned} \tag{3.18}$$

By (3.17) we have

$$\hat{u}(x) \geqq s \quad \text{if and only if } u(x) \geqq s \quad \text{a.e. in } \Omega, \quad \forall s \in \mathbb{R} \setminus]a, b],$$

$$\hat{u}(x) \geqq s \quad \text{if and only if } u(x) \geqq \hat{s} \quad \text{a.e. in } \Omega, \quad \forall s \in]a, b];$$

that is

$$H_s(\hat{u}(x)) = H_s(u(x)) \quad \text{a.e. in } \Omega \quad \forall s \in \mathbb{R} \setminus]a, b],$$

$$H_s(\hat{u}(x)) = H_{\hat{s}}(u(x)) \quad \text{a.e. in } \Omega \quad \forall s \in]a, b],$$

whence

$$B_{\hat{u}}(s) = B_u(s) \quad \forall s \in \mathbb{R} \setminus]a, b[,$$

$$B_{\hat{u}}(s) = B_u(\hat{s}) \quad \forall s \in]a, b[;$$

therefore, also recalling (3.15) and (3.16), we have

$$\tilde{\Psi}_\xi(\hat{u}) - \tilde{\Psi}_\xi(u) = \int_a^b [B_{\hat{u}}(s) - B_u(s)]\, ds = \int_a^b [B_u(\hat{s}) - B_u(s)]\, ds \leqq 0. \tag{3.19}$$

Moreover, as $\tilde{\phi} \leqq \phi$ in $\mathbb{R}$, we have

$$\tilde{\Psi}_\xi(u) \leqq \Psi_\xi(u); \tag{3.20}$$

then by (3.18)-(3.20) we get

$$\Psi_\xi(\hat{u}) \leqq \Psi_\xi(u) = \inf \Psi_\xi;$$

that is, $\xi \in \partial\Psi(\hat{u})$. Thus (3.13) also holds (for $A =]a, b[$).[1]

Now we prove (3.14). If it did not hold, then we would have $\tilde{\Psi}_\xi(u) < \Psi_\xi(u)$. Then by (3.18)-(3.20) we would get $\Psi_\xi(\hat{u}) < \Psi_\xi(u)$; but this would contradict the assumption that $\xi \in \partial\Psi(u)$. So the thesis holds if A is reduced to a single interval. If $A = \cup_{i \in I}]a_i, b_i[$, union of pairwise disjoint intervals, then we set

$$Z_i = \{x \in \Omega : a_i < u(x) < b_i\}, \quad \lambda_1^i := [\phi(b_i) - \phi(a_i)]/(b_i - a_i), \quad \lambda_2^i := \phi(a_i)$$

for any $i \in I$. Note that (cf. (3.15))

$$\begin{aligned} \tilde{\Psi}_\xi(v) = & \int_{\Omega \setminus \cup_{i \in I} Z_i} \phi_{\xi(x)}(v(x))\, dx + \sum_{i \in I} \int_{a_i}^{b_i} B_v(s)\, ds \\ & + \sum_{i \in I} a_i \int_{Z_i} [\lambda_1^i - \xi(x)]\, dx + \sum_{i \in I} \lambda_2^i \mu(Z_i) + \int_{\mathbb{R} \setminus A} \Lambda(H_s(v))\, ds; \end{aligned} \tag{3.21}$$

then by Lemma 2 the previous construction "$u \to \hat{u}$" can be performed simultaneously on all the intervals $]a_i, b_i[$. □

---

[1] By (3.14), $\tilde{\Psi}_\xi(u) = \Psi_\xi(u)$; as $\Psi_\xi(\hat{u}) = \inf \Psi_\xi(u)$, we also have $\Psi_\xi(\hat{u}) = \Psi_\xi(\hat{u})$; hence, by (3.18), $\tilde{\Psi}_\xi(\hat{u}) = \tilde{\Psi}_\xi(u)$. Therefore, by (3.19) the equality holds in (3.16); that is, for no choice of $\hat{s} \in]a, b[$ the inequality can be strict. Thus $B_u(s)$ is constant for $a < s < b$, and $\hat{s}$ can be replaced by any other $s \in]a, b[$.

Note that replacing $\hat{s}$ by another s in (3.17) can yield a function $\hat{u} \neq \hat{u}_{\hat{s}} := (\hat{u})$ only if $\phi(y) = l(y)$ for some $y \in]a, b[$.

This note is based on a remark by M. Paolini.

Now we present a set of sufficient conditions for the existence of an absolute minimum u of Ψ_ξ, or equivalently for $\xi \in \partial\Psi(u)$, to hold for some u.

PROPOSITION 3. *Assume that*

$$\lim_{y\to\pm\infty} \frac{\phi(y)}{|y|} = +\infty, \tag{3.22}$$

$$\Lambda \text{ is strongly lower semicontinuous in } L^1(\Omega), \tag{3.23}$$

$$\text{The injection } \mathrm{Dom}\,(\Lambda) \to L^1(\Omega) \text{ is compact.} \tag{3.24}$$

Then for any $\xi \in L^\infty(\Omega)$ there exists an absolute minimum of Ψ_ξ.

Proof. By (3.22) it is easy to check that there exist $M, N > 0$ such that

$$\phi(y) - \phi(M) \geqq \|\xi\|_{L^\infty(\Omega)}(y - M) \quad \forall y \geqq M,$$

$$\phi(y) - \phi(-N) \geqq -\|\xi\|_{L^\infty(\Omega)}(y + N) \quad \forall y \leqq -N;$$

hence

$$\phi_{\xi(x)}(y) - \phi_{\xi(x)}(M) = \phi(y) - \phi(M) - \xi(x)(y - M) \geqq 0 \quad \text{a.e. in } \Omega \quad \forall y \geqq M, \tag{3.25}$$

$$\phi_{\xi(x)}(y) - \phi_{\xi(x)}(-N) = \phi(y) - \phi(-N) - \xi(x)(y + N) \geqq 0 \quad \text{a.e. in } \Omega \quad \forall y \leqq -N. \tag{3.26}$$

Let $\{u_n\}$ be a minimizing sequence of Ψ_ξ, and set

$$\tilde{u}_n(x) := \min\{M, \max[u(x), -N]\} \quad \text{a.e. in } \Omega.$$

By (3.25), (3.26), and (2.5), we have

$$\phi_{\xi(x)}(\tilde{u}_n(x)) \leqq \phi_{\xi(x)}(u_n(x)) \quad \text{a.e. in } \Omega,$$

$$\Lambda(\tilde{u}_n) \leqq \Lambda(u_n);$$

hence also $\{\tilde{u}_n\}$ is a minimizing sequence; moreover,

$$\|\phi_{\xi(x)}(\tilde{u}_n(x))\|_{L^\infty(\Omega)}, \Lambda(\tilde{u}_n) \leqq \text{constant}.$$

Hence by (3.24) there exists $u \in L^1(\Omega)$ such that, possibly extracting a subsequence, we get

$$\tilde{u}_n \to u \quad \text{strongly in } L^1(\Omega) \quad \text{a.e. in } \Omega; \tag{3.27}$$

hence, by the dominated convergence theorem and by (3.23),

$$\int_\Omega \phi_{\xi(x)}(\tilde{u}_n(x))\,dx \to \int_\Omega \phi_{\xi(x)}(u(x))\,dx,$$

$$\liminf \Lambda(\tilde{u}_n) \geqq \Lambda(u).$$

Thus $\Psi_\xi(u) = \inf \Psi_\xi$. □

Remarks. (i) By Proposition 2 in § 2, the functionals V, Λ_r, and $\tilde{\Lambda}_r$, with $0 < r < 1$, fulfill (3.23) and, if Ω is bounded and smooth, also (3.24).

(ii) If $\varphi(y)/y$ is bounded as $|y| \to +\infty$, then Ψ_ξ may have no absolute minimum, as is easy to check.

First applications of Theorem 1. (i) Let us take

$$\phi(y) = (y^2 - 1)^+ \quad \forall y \in \mathbb{R}.$$

For any $\xi \in \mathbb{R}$, $\phi_\xi : \mathbb{R} \to \mathbb{R} : y \mapsto \phi(y) - \xi y$ has a minimum $\tilde{y} \in \mathbb{R}\setminus]-1, 1[$; otherwise stated, the restriction of $\phi'(y) = 2(y-1)^+ - 2(y+1)^-$ to $\mathbb{R}\setminus]-1, 1[$ is surjective, even if its graph is not maximal monotone.

Something similar occurs for space distributed systems. Let $\Lambda \in GC(\Omega)$ be such that (3.23) and (3.24) hold; then by Proposition 1, for any $\xi \in L^\infty(\Omega)$, Ψ_ξ has a minimum, that is, $\partial\Psi$ is surjective. Hence, by part (i) of Theorem 1, the restriction of $\partial\Psi$ to $\{v \in L^1(\Omega) : v(x) \notin]-1, 1[\text{ a.e. in } \Omega\}$ is also onto $L^\infty(\Omega)$.

(ii) If we take

$$\phi(y) := \begin{cases} \frac{a}{2}(1-y^2) & \text{if } |y| \leqq 1 \\ +\infty & \text{if } |y| > 1 \end{cases} \qquad (a: \text{constant} > 0)$$

(cf. Fig. 2), then

$$\partial\phi(y) = \begin{cases} \mathbb{R}^- & \text{if } y = -1, \\ \varnothing & \text{if } y \in \mathbb{R}\setminus\{-1, 1\} \\ \mathbb{R}^+ & \text{if } y = 1; \end{cases}$$

therefore $\partial\phi|_{\{-1,1\}}$ is surjective. Then, by (i) of Theorem 1, for any $\Lambda \in GC(\Omega)$ such that (3.23) and (3.24) hold, the restriction of $\partial\Psi$ to $\{v \in L^1(\Omega) : v(x) = \pm 1 \text{ a.e. in } \Omega\}$ is onto $L^\infty(\Omega)$.

LEMMA 3 [4, p. 20]. *Let V be a separated locally convex vector space and $F : V \to \mathbb{R} \cup \{+\infty\}$. So for any $u \in V$, if $\partial F(u) \neq \varnothing$, then $F(u) = F^{**}(u)$; if $F(u) = F^{**}(u)$, then $\partial F(u) = \partial F^{**}(u)$.*

THEOREM 2. *Let $\Lambda \in GC(\Omega)$ (cf. § 2) and ϕ fulfill* (3.1) *and* (3.2); *set* (3.3) *and* (3.4). *Assume that*

(3.28) *All the connected components of $S := \{y \in \mathbb{R} : \phi(y)^{**} < \phi(y)\}$ are bounded.*

For any $u \in L^1(\Omega)$, if

$$\partial\Psi(u) \neq \varnothing \quad \text{in } L^\infty(\Omega), \tag{3.29}$$

then

$$\Phi(u) = \Phi^{**}(u), \tag{3.30}$$

$$\partial\Phi(u) \neq \varnothing \quad \text{in } L^\infty(\Omega), \tag{3.31}$$

$$\exists \eta \in \partial\Phi(u) : \forall \xi \in \partial\Psi(u), \qquad \|\eta\|_{L^\infty(\Omega)} \leqq \|\xi\|_{L^\infty(\Omega)}. \tag{3.32}$$

Remark. By Lemma 3, (3.30) and (3.31) yield

$$\partial\Phi(u) = \partial\Phi^{**}(u) \neq \varnothing \quad \text{in } L^\infty(\Omega); \tag{3.33}$$

(3.30) *and* (3.33) *are, respectively, equivalent to*

$$\phi(u) = \phi^{**}(u), \quad \partial\phi(u) = \partial\phi^{**}(u) \neq \varnothing \quad \text{a.e. in } \Omega. \tag{3.34}$$

Proof. The set S is open, since ϕ is lower semicontinuous and ϕ^{**} is affine in each connected component of S. Hence, by (3.28), S is a countable union of disjoint *bounded* open intervals. Then, by (ii) of Theorem 1, $\mu(\{x \in \Omega : u(x) \in S\}) = 0$; namely,

$$\phi(u) = \phi^{**}(u) \quad \text{a.e. in } \Omega; \tag{3.35}$$

therefore (3.30) holds.

To prove (3.31) and (3.32), we fix any $\xi \in \partial\Psi(u)$; we also set $N := \|\xi\|_{L^\infty(\Omega)}$ and for any $M \in \mathbb{N}$

$$\Omega_M := \{x \in \Omega : u(x) \leqq -M \text{ or } \exists \eta \in \partial^{**}(u(x)) : \eta \leqq N\},$$

$$y_M := \sup\,(]-\infty, -M] \cup \{y \in \mathbb{R} : \exists \eta \in \partial\phi^{**}(y) : \eta \leqq N\}),$$

$$\tilde{u}(x) := \begin{cases} u(x) & \text{in } \Omega_M, \\ y_M (< u(x)) & \text{in } \Omega\setminus\Omega_M. \end{cases}$$

Then

$$\varphi_{\xi(x)}(u(x))-\varphi_{\xi(x)}(\tilde{u}(x))\begin{cases}=0 & \text{a.e. in } \Omega,\\ \geqq \phi(u(x))-\phi(y_M)-N(u(x)-y_M) & \text{a.e. in } \Omega\backslash\Omega_M,\end{cases}$$

$$\Phi_\xi(u)\geqq\Phi_\xi(\tilde{u}), \tag{3.36}$$

and the equality holds only if $\mu(\Omega\backslash\Omega_M)=0$. Moreover, by (2.5)

$$\Lambda(\tilde{u})\leqq\Lambda(u);$$

hence $\Psi_\xi(\tilde{u})\leqq\Psi_\xi(u)$. On the other hand, $\Psi_\xi(u)=\inf\Psi_\xi$, as $\xi\in\partial\Psi(u)$; hence $\Psi_\xi(\tilde{u})=\Psi_\xi(u)$ and the equality holds also in (3.36), whence

$$\mu(\Omega\backslash\Omega_M)=0 \quad \forall M\in\mathbb{N}.$$

Note that by Lemma 3 and (3.35), $\partial\phi(u(x))=\partial\phi^{**}(u(x))$ almost everywhere in Ω; then

$$\begin{aligned}&\mu(\{x\in\Omega:\partial\phi(u(x))=\varnothing \text{ or } \forall\eta\in\partial\phi(u(x)),\ \eta>N\})\\ &\quad=\mu(\{x\in\Omega:\partial\phi^{**}(u(x))=\varnothing \text{ or } \forall\eta\in\partial\phi^{**}(u(x)),\ \eta>N\})\\ &\quad=\mu\Big(\bigcup_{M\in\mathbb{N}}(\Omega\backslash\Omega_M)\Big)=0.\end{aligned}$$

Similarly, we might show that

$$\mu(\{x\in\Omega:\partial\phi(u(x))=\varnothing \text{ or } \forall\eta\in\partial\phi(u(x)),\ \eta<-N\})=0.$$

Therefore (3.31) holds and

$$\mu(\{x\in\Omega:\forall\eta\in\partial\phi(u(x)),\ |\eta|>N\})=0;$$

that is,

$$\forall\xi\in\partial\Psi(u) \quad \exists\eta\in\partial\Phi(u):\|\eta\|_{L^\infty(\Omega)}\leqq\|\xi\|_{L^\infty(\Omega)}.$$

For any $u\in L^1(\Omega)$ such that $\partial\Psi(u)\neq\varnothing$, taking η equal to the element of $\partial\Phi(u)=\partial\Phi^{**}(u)$ of minimum norm, we get (3.32). □

Remark. Equation (3.22) is a sufficient condition for (3.28) to hold. If, for instance, S has a connected component of the form $]a,+\infty[$ $(a\in\mathbb{R})$, then ϕ^{**} is linear in a neighbourhood of $+\infty$ and

$$\lim_{y\to+\infty}\phi(y)/y=\lim_{y\to+\infty}\phi^{**}(y)/y<+\infty.$$

In that case, as remarked after Proposition 3, there exists a $\xi\in L^\infty(\Omega)$ such that the problem of minimizing Ψ_ξ has no solution; that is, $\partial\Psi$ is not surjective.

LEMMA 4 [4, p. 26]. *Let V be a separated locally convex vector space. Let $F_1, F_2: V\to\mathbb{R}\cup\{+\infty\}$ be proper, convex, and lower semicontinuous; let F_1 be continuous at some point of* $\mathrm{Dom}\,(F_1)\cap\mathrm{Dom}\,(F_2)$. *Then for any $u\in V$*

$$\partial(F_1+F_2)(u)=\partial F_1(u)+\partial F_2(u).$$

THEOREM 3. *Let $\Lambda\in GC(\Omega)$ (cf. § 2) and ϕ fulfill* (3.1) *and* (3.2); *set* (3.3) *and* (3.4). *Assume that* (3.28) *holds and that Λ is lower semicontinuous and convex (i.e., $\Lambda^{**}=\Lambda$). Then for any $u\in L^1(\Omega)$*

$$(\Phi+\Lambda)^{**}(u)=\Phi^{**}(u)+\Lambda(u), \tag{3.37}$$

$$\partial(\Phi+\Lambda)(u)=\partial\Phi(u)+\partial\Lambda(u) \quad \text{in } L^\infty(\Omega) \tag{3.38}$$

(here we set $A+\varnothing=\varnothing$ for any $A\subset L^\infty(\Omega)$).

Remarks. (i) If $\partial(\Phi+\Lambda)(u)\neq\varnothing$, then by (3.38)

$$\partial\Lambda(u)\neq\varnothing \quad \text{in } L^\infty(\Omega);$$

that is,

$$\exists\eta\in L^\infty(\Omega):\quad \forall v\in \mathrm{Dom}\,(\Lambda)\quad \int_\Omega \eta(u-v)\,dx\geqq\Lambda(u)-\Lambda(v); \tag{3.39}$$

this can be regarded as a regularity result.

(ii) By (3.37) and (3.38), for any $u\in L^1(\Omega)$ we have

$$\partial(\Phi+\Lambda)^{**}(u)=\partial(\Phi^{**}+\Lambda)(u)=\partial\Phi^{**}(u)+\partial\Lambda(u). \tag{3.40}$$

Proof. (i) For any $u\in L^1(\Omega)$, we have

$$\bar\Psi(u):=(\Phi^{**}+\Lambda)(u)\leqq(\Phi+\Lambda)(u)=:\Psi(u).$$

Hence to prove that $\bar\Psi$ is actually the lower semicontinuous convex hull of Ψ, it is sufficient to show that for any $u\in L^1(\Omega)$ such that $\partial\bar\Psi(u)\neq\varnothing$ and for any $\xi\in\partial\bar\Psi(u)$, there exists a $\hat u\in L^1(\Omega)$ such that

$$\Psi_\xi(\hat u)=\bar\Psi_\xi(u). \tag{3.41}$$

Indeed, by Theorem 1, here applied with Ψ replaced by $\bar\Psi$, there exists a $\hat u\in L^1(\Omega)$ such that

$$\bar\Psi_\xi(\hat u)=\bar\Psi_\xi(u) \tag{3.42}$$

and $\Phi(\hat u)=\Phi^{**}(\hat u)$; that is, $\Psi_\xi(\hat u)=\bar\Psi_\xi(\hat u)$; then (3.41) holds.

(ii) In (3.38) the inclusion $\supset$ is obvious; now we will show the opposite. Let us fix any $u\in L^1(\Omega)$ such that $\partial(\Phi+\Lambda)(u)\neq\varnothing$ and any $\xi\in\partial(\Phi+\Lambda)(u)$. Let $\hat\phi$ be the largest Lipschitz continuous, convex function $\mathbb{R}\to\mathbb{R}$ such that

$$\hat\phi(y)\leqq\phi(y),\quad |\hat\phi'(y)|\leqq N:=\|\xi\|_{L^\infty(\Omega)}\quad \text{a.e. in } \mathbb{R}.$$

The set $T:=\{y\in\mathbb{R}:\hat\phi(y)=\phi^{**}(y)\}$ is a closed interval with extremes a, b, with $-\infty\leqq a<b\leqq+\infty$. We can exclude the trivial case in which $a=b$; moreover,

$$\text{If } a>-\infty, \text{ then } \hat\phi(y)=-Ny\quad \forall y\leqq a,$$

$$\text{If } b<+\infty, \text{ then } \hat\phi(y)=Ny\quad \forall y\geqq b.$$

For any $y\in\mathbb{R}\setminus T$ such that $\partial\phi(y)\neq\varnothing$ and any $\eta\in\partial\phi(y)$, we have $|\eta|>N$; then by (3.32) and (3.34) we get

$$\hat\phi(u)=\phi^{**}(u)=\phi(u)\quad \text{a.e. in } \Omega. \tag{3.43}$$

We set

$$\hat\Phi(v):=\int_\Omega\hat\phi(v(x))\,dx\quad \forall v\in L^1(\Omega),\quad \hat\Psi:=\hat\Phi+\Lambda,\quad \hat\Psi_\xi:=\hat\Phi_\xi+\Lambda;$$

then by (3.43) we have

$$\begin{aligned}&\hat\Phi(u)=\Phi^{**}(u)=\Phi(u),\qquad \hat\Psi(u)=\Psi^{**}(u)=\Psi(u),\\ &\partial\hat\Phi(u)\subset\partial\Phi^{**}(u)=\partial\Phi(u).\end{aligned} \tag{3.44}$$

(iii) Now we want to show that

$$\xi\in\partial(\hat\Phi+\Lambda)(u). \tag{3.45}$$

To this aim, for any $w \in L^1(\Omega)$ we set

$$\tilde{w}(x) := \min\{a, \max[w(x), b]\} \quad \text{a.e. in } \Omega;$$

then

$$\hat{\Phi}(\tilde{w}) = \Phi^{**}(\tilde{w})$$

and by (2.5)

$$\Lambda(\tilde{w}) \leqq \Lambda(w);$$

hence

$$\hat{\Psi}_\xi(\tilde{w}) - \hat{\Psi}_\xi(w) = \hat{\Phi}(\tilde{w}) - \hat{\Phi}(w) + \Lambda(\tilde{w}) - \Lambda(w) - \int_\Omega \xi(\tilde{w} - w)\, dx$$

$$\leqq -N \int_\Omega |\tilde{w} - w|\, dx - \int_\Omega \xi \cdot (\tilde{w} - w)\, dx \leqq 0.$$

Then, still for any $w \in L^1(\Omega)$, by (3.44) we have

$$\hat{\Psi}_\xi(w) \geqq \hat{\Psi}_\xi(\tilde{w}) = \Psi_\xi^{**}(\tilde{w}) \geqq \inf \Psi_\xi^{**} = \inf \Psi_\xi = \Psi_\xi(u) = \hat{\Psi}_\xi(u);$$

thus $\xi \in \partial\hat{\Psi}(u)$. So we have shown (3.45).

(iv) By (3.45), since $\xi \in \partial(\Phi + \Lambda)(u)$ is arbitrary we have

$$\partial(\Phi + \Lambda)(u) \subset \partial(\hat{\Phi} + \Lambda)(u); \tag{3.46}$$

moreover, by Lemma 4 and (3.44), we have

$$\partial(\hat{\Phi} + \Lambda)(u) = \partial\hat{\Phi}(u) + \partial\Lambda(u) \subset \partial\Phi(u) + \partial\Lambda(u). \tag{3.47}$$

Finally, (3.46) and (3.47) yield

$$\partial(\Phi + \Lambda)(u) \subset \partial\Phi(u) + \partial\Lambda(u). \qquad \square$$

DEFINITION. For any Banach space B, any $u \in B$, and any function $F: B \to \mathbb{R} \cup \{+\infty\}$, we define the *local subdifferential* $\partial_{\text{loc}} F(u)$ as follows:

$$\partial_{\text{loc}} F(u) := \{\xi \in B' : \exists \varepsilon > 0 : \forall v \in B, \text{ if } \|u - v\| \leqq \varepsilon \text{ then } F(u) - F(v) \leqq \langle \xi, u - v \rangle\}. \tag{3.48}$$

Note that

(3.49) $\xi \in \partial_{\text{loc}} F(u)$ if and only if u is either a relative or an absolute minimum of $F_\xi : B \to \mathbb{R} \cup \{+\infty\} : v \mapsto F(v) - \langle \xi, v \rangle$.

This statement can be compared with the obvious property that $\xi \in \partial F(u)$ if and only if u is an absolute minimum of F_ξ.

THEOREM 4. *Let* $\Lambda \in GC(\Omega)$ (*cf.* § 2) *and* ϕ *fulfill* (3.1) *and* (3.2); *set* (3.3) *and* (3.4). *For any* $u \in L^1(\Omega)$, *if*

$$\partial_{\text{loc}} \Psi(u) \neq \varnothing \quad \text{in } L^\infty(\Omega), \tag{3.50}$$

then

$$\partial_{\text{loc}} \phi(u(x)) \neq \varnothing \quad \text{a.e. in } \Omega. \tag{3.51}$$

Remark. Inequality (3.51) can be compared with (3.31); however, (3.51) is not equivalent to $\partial_{\text{loc}} \Phi(u) \neq 0$. In fact, $\partial_{\text{loc}} \Phi(u) = \partial\Phi(u)$ in $L^\infty(\Omega)$ (cf. Proposition 5 in § 4), although in general $\partial_{\text{loc}} \phi(v) \neq \partial\phi(v)$ in $\mathbb{R}$.

Proof. By assumption there exist $\xi \in L^\infty(\Omega)$ and $\varepsilon > 0$ such that $\Psi_\xi(u) \leqq \Psi_\xi(v)$ for any $v \in L^1(\Omega)$ with $\|u - v\|_{L^1(\Omega)} \leqq \varepsilon$. Let us fix any $a, b \in \mathbb{R}$ with $0 < b - a < \varepsilon/\mu(\Omega)$; then u is an absolute minimum of Ψ_ξ restricted to

$$Y^u_{a,b} := \{v \in L^1(\Omega) : v(x) = u(x) \text{ in } \{u \leqq a\} \cup \{u \geqq b\},\ a \leqq v(x) \leqq b \text{ in } \{a \leqq u \leqq b\}\}$$

(here we set, e.g., $\{u \leqq a\} := \{x \in \Omega : u(x) \leqq a\}$); in fact, for any $v \in Y^u_{a,b}$, $\|u - v\|_{L^1(\Omega)} \leqq (b-a) \cdot \mu(\Omega) \leqq \varepsilon$. We denote the indicator function of $]a, b[$ by $I_{]a,b[}$; that is,

$$I_{]a,b[}(y) := \begin{cases} 0 & \text{if } a < y < b, \\ +\infty & \text{otherwise,} \end{cases}$$

and set

$$S_{a,b} := \{y \in]a, b[: (\phi + I_{]a,b[})^{**}(y) < \phi(y)\};$$

as shown for S in the proof of Theorem 3, $S_{a,b}$ is also a countable union of disjoint open intervals.

We are then reduced to showing that $\mu(\{x \in \Omega : u(x) \in S_{a,b}\}) = 0$; the latter can be proved by the argument used for (3.14) in Theorem 1. □

A question. Does (3.50) entail $\partial\Lambda(u) \neq \varnothing$ in $L^\infty(\Omega)$?

4. Multiphase systems.

4.1. Stable and metastable states. In several models the stationary configurations of a system correspond to either the relative or the absolute minima of a potential functional; for multiphase systems this potential is nonconvex. Here we will consider systems characterized by a *scalar* state variable. We will distinguish between *distributed* and *nondistributed* systems, namely, between systems with and without space dependence. In a scalar nondistributed system, the state is characterized by a a real variable, and as a potential we can consider a (possibly nonconvex) lower semicontinuous function $\phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$. A "double-well" potential ϕ as sketched in Fig. 1 corresponds to a two-phase system. A "triple-well" potential represents a three-phase system, and so on. Certain two-phase systems can be represented by the potential ϕ sketched in Fig. 2 (see § 5).

Any state that can be attained by a system for arbitrarily long times is said to have a *stable equilibrium*; any state that will eventually decay into another is said to have a *metastable equilibrium.* An example of the latter is given by a supercooled liquid (water below 0°C, e.g.) (cf. § 5 and [3, Chap. 3]).

If the potential has a unique absolute minimum, this corresponds to a stable state. If it also has a relative (nonabsolute) minimum, this represents a metastable state. (To be more precise, the latter statement requires that lower potential states be accessible from the considered relative minimum state—a condition we will always assume and that will hold in the examples we will present.)

Metastable states decay because of thermodynamic fluctuations that allow the system to explore nearby states. Although fluctuations are a stochastic phenomenon, at least in some cases, there is a net separation between states whose persistence for a given time period is almost sure, and others whose persistence is almost impossible (cf. [3, Chap. 3]). So a deterministic model is not a priori excluded.

As an example, let us consider a nondistributed system associated to the double-well potential ϕ sketched in Fig. 1. Here the absolute minimum $u = h$ corresponds to a stable state; the relative minimum $u = b$ represents a metastable state, which persists as long as the thermodynamic fluctuations do not bring the system from $u = b$ to a state $u \geqq d$, from which the system would evolve towards the state $u = h$.

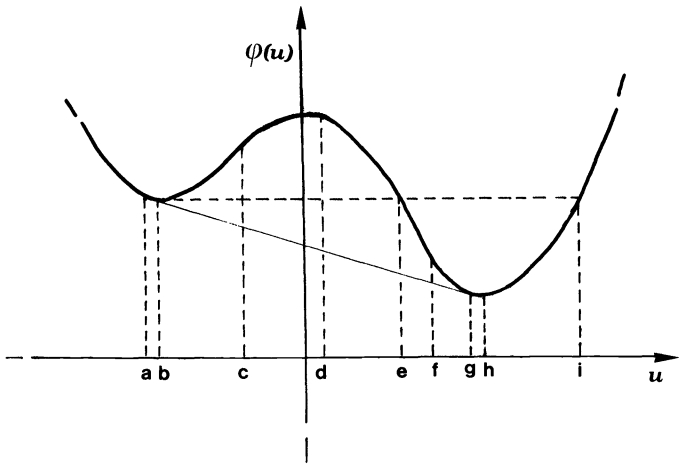


FIG. 1. *Double-well potential function* $\phi \in C^1(\mathbb{R})$. $u = b$ *is a relative minimum*; $u = h$ *is an absolute minimum*; $u = c$ *and* $u = f$ *are flexi*; $u = d$ *is a relative maximum*; $\phi(b) = \phi(e) = \phi(i)$. *The lower semicontinuous, convex regularized* ϕ^{**} *of* ϕ *coincides with* ϕ *in* $\mathbb{R}\backslash]a, g[$ (*the drawn segment is tangent to the graph of* ϕ *at* $u = a$, $u = g$). *So,*

$$\partial\varphi(u) = \partial_{\text{loc}}\varphi(u) \neq \varnothing \qquad \forall u \in \mathbb{R}\backslash]a, g[,$$

$$\partial\varphi(u) = \varnothing, \quad \partial_{\text{loc}}\varphi(u) \neq \varnothing \qquad \forall u \in [a, g]\backslash [c, f],$$

$$\partial\varphi(u) = \partial_{\text{loc}}\varphi(u) = \varnothing \qquad \forall u \in [c, f].$$

In a time period of length T the probability of the occurrence of a fluctuation from a state u_1 to another state u_2 ($> u_1$, say) decreases as the potential variation $\phi(u_2) - \phi(u_1)$ increases if $\phi(\xi)$ is comprised between $\phi(u_1)$ and $\phi(u_2)$ for $\xi \in [u_1, u_2]$. More precisely, the following occurs for liquid-solid transformations, and seems to be a typical picture of phase transitions in general. There is a critical value $\Xi(T) > 0$ such that for any $\delta > 0$ the occurrence of fluctuations with potential variations equal to δ is extremely likely (extremely unlikely, respectively) if $\delta < \Xi(T)$ ($\delta > \Xi(T)$, respectively) [3, p. 70].

If the potential has more than one absolute minimum, then the state will oscillate among the absolute minima that are separated by a "potential barrier" smaller than $\Xi(T)$. We can imagine that the stability is divided among these equivalent states. A similar effect occurs for relative minima that correspond to a same value of the potential and are also separated by potential barriers smaller than $\Xi(T)$.

However, the system cannot oscillate between two states corresponding to different values of the potential, since the probability of going from the state with the larger potential to that with the smaller is much greater than the probability of the opposite fluctuation.

We summarize the previous discussion in the following statement.

PROPOSITION 4. *The absolute minima of the potential correspond to stable equilibrium states. The relative minima represent metastable equilibrium states, namely, states that can persist for some time, but that will eventually decay. For each of the latter states, the duration of its persistence depends on the depth of the "potential well" in which the relative minimum stays.*

In this section we will study only the properties of absolute and relative minima of a given potential. Thus we will describe stable and metastable states of the corresponding system, without attempting to characterize the metastable states that can

persist for a given time period. The latter question will be considered in § 5; there, dealing with a solid-liquid system, we will introduce a modified potential for which the following occurs: The absolute minima of the original and modified potentials coincide, but the relative minima of the modified potential are just a subset of those of the original one, and correspond to the metastable states that persist for the given time period.

In several cases the potential functional contains a linear term, which can be controlled; this allows us to explore all the possible configurations of the system. For instance, in solid-liquid systems this linear term is proportional to the relative temperature (cf. § 5).

4.2. Distributed systems. For a scalar distributed system occupying a bounded Euclidean domain Ω, the state is represented by a real function $u:\Omega\to\mathbb{R}$. We assume that (3.1) and (3.2) hold, so that (3.3) is also fulfilled. We then fix any $\xi\in L^\infty(\Omega)$ and consider a potential function of the form

$$\phi_{\xi(x)}(v) := \phi(v) - \xi(x)v \quad \forall v\in\mathbb{R} \quad \text{a.e. in } \Omega, \tag{4.1}$$

to which we associate the potential functional

$$\Phi_\xi(v) := \int_\Omega \phi_{\xi(x)}(v(x))\,dx \quad \forall v\in L^1(\Omega). \tag{4.2}$$

PROPOSITION 5. *For any $\xi\in L^\infty(\Omega)$, $u\in L^1(\Omega)$ is an absolute minimum of Φ_ξ in $L^1(\Omega)$ if and only if, almost everywhere in Ω, $u(x)$ is an absolute minimum of $\phi_{\xi(x)}$ in $\mathbb{R}$.*

Φ_ξ has no relative (nonabsolute) minimum with respect to the strong topology of $L^1(\Omega)$.

Proof. The first part is obvious. To prove the second statement, first, for any $R\subset\mathbb{R}$ and any $y\in\mathbb{R}$ we set

$$d(y,R) := \inf_{r\in R} |y-r|,$$

$$\operatorname{proj}(y,R) := \{r\in R : |y-r| = d(y,R)\}.$$

Almost everywhere in Ω, let $M(x)\subset\mathbb{R}$ be the set of the absolute minima of $\phi_{\xi(x)}$. By contradiction let $u\in L^1(\Omega)$ be a relative (nonabsolute) minimum of Φ_ξ; then by the first part of this proposition there exists a measurable set $A\subset\Omega$ such that $\mu(A)>0$ and

$$u(x)\notin M(x) \quad \text{a.e. in } A.$$

For any $\varepsilon>0$, let us consider a measurable set $A_\varepsilon\subset A$ such that $\mu(A_\varepsilon)>0$ and $\int_{A_\varepsilon} d(u(x),M(x))\,dx \leqq \varepsilon$, and let

$$u_\varepsilon(x)\begin{cases} = u(x) & \text{if } x\in\Omega\setminus A_\varepsilon,\\ \in \operatorname{proj}(u(x),M(x)) & \text{if } x\in A_\varepsilon;\end{cases}$$

then

$$\Phi_\xi(u_\varepsilon) < \Phi_\xi(u) \quad \text{and} \quad \|u-u_\varepsilon\|_{L^1(\Omega)} \leqq \varepsilon,$$

contradicting the assumption on u. □

By the first part of the previous proposition, as ξ spans $L^\infty(\Omega)$, the set of all the absolute minima of Φ_ξ defines a *multiphase structure* in the following sense.

DEFINITION. For any family $\mathcal{F}$ of states $u:\Omega\to\mathbb{R}$, the set $P := \cup_{u\in\mathcal{F}} u(\Omega)$ $(\subset\mathbb{R})$ is called *phase structure of $\mathcal{F}$*. The family $\mathcal{F}$ is said to have an *m-phase structure* if P has m connected components. For any $u\in\mathcal{F}$ and any connected component C of P, $u^{-1}(C)$ will be called a *phase*.

By Propositions 4 and 5, a system represented by the potential Φ_ξ can have stable states but admits no metastable states. To account for the latter we will introduce a space interaction term.

Space interaction terms. We will use no term of the form $\Lambda^p(u) := \int_\Omega |\nabla u(x)|^p \, dx$, $1 \leqq p \leqq \infty$; in fact, by well-known trace theorems in Sobolev spaces [1], the condition $\Lambda^p(u) < +\infty$ is not consistent with the discontinuities along surfaces that occur in multiphase states. Rather, we will consider the functionals V, Λ_r, and $\tilde{\Lambda}_r$ $(0 < r < 1)$, defined in § 2; they, and more generally any functional $\Lambda \in GC(\Omega)$ (cf. § 2), are compatible with a *multiphase structure.* The physical meaning of these functionals will be discussed in § 5.

Thus, if ϕ represents the potential density function, as a global potential functional we consider $\Psi := \Phi + \Lambda$, or also, including linear perturbations, $\Psi_\xi := \Phi_\xi + \Lambda$, for any $\xi \in L^\infty(\Omega)$. As we have seen (cf. Proposition 3 of § 3) under suitable assumptions Ψ_ξ has an absolute minimum. Ψ_ξ can also have a *local* relative (nonabsolute) minimum u. By this we mean that

$$\exists \delta > 0 : \forall w \in L^1(\Omega) \text{ such that } u - w \text{ has compact support in } \Omega \text{ and } \|u - w\|_{L^1(\Omega)} \leqq \delta,\ (\Phi_\xi + \Lambda)(u) \leqq (\Phi_\xi + \Lambda)(w), \tag{4.3}$$

and that u is not an absolute minimum. Note that the class of local relative minima includes that of relative minima.

We will check the existence of such a u in the case of $\Lambda = V$, ϕ as in Fig. 1, and $\xi \equiv 0$ almost everywhere in Ω. Here $u \equiv b$ in Ω is a local relative (nonabsolute) minimum, since any small local variation of u in $L^1(\Omega)$ increases the V-term more than it can decrease the Φ-term. Let us consider any function $w \in L^1(\Omega)$ such that $u - w$ has compact support in Ω and $\|u - w\|_{L^1(\Omega)}$ is "small." It is clear that taking values where ϕ is larger than $\phi(b)$ does not help to reduce $\Phi + V$; hence we can assume that there exists a set $S \subset \Omega$ such that

$$w = b \quad \text{a.e. in } \Omega \backslash S, \qquad e \leqq w \leqq i \quad \text{a.e. in } S.$$

To reduce $V(w)$ it is "convenient" that S be a sphere, of radius R, say. Then, for instance, for $\Omega \subset \mathbb{R}^3$,

$$V(w) - V(u) = V(w) \geqq (e - b) 4\pi R^2,$$

$$\Phi(w) - \Phi(u) \geqq [\phi(h) - \phi(b)] 4\pi R^3/3;$$

hence for R sufficiently small, $(\Phi + V)(w) > (\Phi + V)(u)$. Therefore u is a local relative (nonabsolute) minimum of $\Phi + V$. A similar argument can be used for $\Lambda = \Lambda_r$, for any $r \in]0, 1[$.

Remark. The distinction between local and global relative minima of the free enthalpy is related to the phenomena of *homogeneous* and *heterogeneous nucleation* [3, Chap. 3].

4.3. Applications of Theorems 2 and 4 (cf. § 3). For any $\xi \in L^\infty(\Omega)$, obviously $u \in L^1(\Omega)$ is an absolute minimum of Ψ_ξ if and only if $\xi \in \partial\Psi(u)$. By Theorem 2 of § 3 this entails that $\partial\Phi(u) \neq \varnothing$, or equivalently

$$\partial\phi(u(x)) \neq \varnothing \quad \text{a.e. in } \Omega; \tag{4.4}$$

hence u can be a multiphase state. For instance, for ϕ as in Fig. 1,

$$\{y \in \mathbb{R} : \partial\phi(y) \neq \varnothing\} = \mathbb{R} \backslash]a, g[. \tag{4.5}$$

Note that in general $\partial\Psi(u) \neq \partial\Phi(u)$; moreover, for any $\xi \in L^\infty(\Omega)$ the absolute minima of Ψ_ξ belong to Dom (Λ), and hence they do not generally coincide with the absolute minima of Φ_ξ.

Thus we have the following result.

PROPOSITION 6. *For any $\xi \in L^\infty(\Omega)$, if u is an absolute minimum of Ψ_ξ in $L^1(\Omega)$, then* (4.4) *holds.*

Thus the introduction of the functional $\Lambda = V$, or $\Lambda = \Lambda_r$, or $\Lambda = \tilde{\Lambda}_r$ $(0 < r < 1)$ (cf. § 2) *may modify the absolute minima of Φ_ξ, but it maintains their phase structure.*

Let us now consider *local relative* minima. Let us fix any $\xi \in L^\infty(\Omega)$. The argument of Proposition 5 can be used to show that Φ_ξ has no local relative (nonabsolute) minimum; instead $\Psi_\xi := \Phi_\xi + \Lambda$ may have such a minimum. Any $u \in L^1(\Omega)$ is either a relative or an absolute minimum of Ψ_ξ if and only if $\xi \in \partial_{\mathrm{loc}}\Psi(u)$ (defined in (3.48)); by Theorem 4 of § 3, this entails that

$$\partial_{\mathrm{loc}}\phi(u(x)) \neq \varnothing \quad \text{a.e. in } \Omega. \tag{4.6}$$

Thus also the relative minima have a *phase structure.* The latter can be different from that of the absolute minima of Ψ_ξ. For instance, for ϕ as in Fig. 1,

$$\{y \in R\colon \partial_{\mathrm{loc}}\phi(y) \neq \varnothing\} = \mathbb{R}\backslash[c, f] \neq \{y \in R\colon \partial\phi(y) \neq \varnothing\}; \tag{4.7}$$

however, for ϕ as in Fig. 2,

$$\{y \in \mathbb{R}\colon \partial\phi(y) \neq \varnothing\} = \{y \in \mathbb{R}\colon \partial_{\mathrm{loc}}\phi(y) \neq \varnothing\} = \{-1, 1\}. \tag{4.8}$$

We have the following result.

PROPOSITION 7. *The addition of either $\Lambda = V$ or $\Lambda = \Lambda_r$ or $\Lambda = \tilde{\Lambda}_r$ $(0 < r < 1)$ to the nonconvex functional Φ_ξ allows local relative (nonabsolute) minima to occur for suitable ϕ and ξ. For any such Λ and any $\xi \in L^\infty(\Omega)$, if u is a local relative minimum of Ψ_ξ in $L^1(\Omega)$, then* (4.6) *holds. The phase structure of the (local) relative minima of Ψ_ξ can be different from that of its absolute minima.*

Propositions 4, 6, and 7 yield the following conclusions.

PROPOSITION 8. *For any $\xi \in L^\infty(\Omega)$, for a system governed by the potential $\Psi_\xi := \Phi_\xi + \Lambda$, with either $\Lambda = V$ or $\Lambda = \Lambda_r$ or $\Lambda = \tilde{\Lambda}_r$ $(0 < r < 1)$, stable and metastable states have the phase structures* (4.4) *and* (4.6) *respectively; namely,* (4.5) *and* (4.7), *for ϕ as in Fig.* 1.

5. Surface tension effects.

5.1. Supercooling and superheating. First we review the developments of [20]. Let us assume that Ω is occupied by a substance capable of attaining two states, solid and liquid for instance, here represented by $u = -1$ and $u = 1$, respectively. At constant pressure, the potential to be minimized is the total free enthalpy. First we consider the function $\phi_{\xi(x)}(u) := \phi(u) - \xi(x)u$, where $\phi := I_{\{-1,1\}}$ (i.e., $\phi(y) = 0$ if $|y| = 1$, $\phi(y) = +\infty$ if $|y| \neq 1$) and $\xi(x)$ is (proportional to) the relative temperature, namely, the difference between the actual temperature and its equilibrium value for a flat solid-liquid interface. Here, dealing with the stationary problem, we will assume that $\xi(x)$ is prescribed almost everywhere in Ω.

Obviously, for any $\xi \in L^\infty(\Omega)$, $u \in L^1(\Omega)$ is an absolute minimum of Φ_ξ (cf. (4.2)) if and only if

$$u(x) = \begin{cases} -1 & \text{if } \xi(x) < 0, \\ 1 & \text{if } \xi(x) > 0 \end{cases} \tag{5.1}$$

almost everywhere in Ω. At the interface $\mathcal{S}$ between the phases, this yields

$$\xi = 0 \quad \text{on } \mathcal{S}; \tag{5.2}$$

more precisely, the latter condition holds only at points of continuity of ξ on $\mathcal{S}$.

The functional $\Phi_\xi : L^1(\Omega) \to \mathbb{R} \cup \{+\infty\}$ has no relative (nonabsolute) minima, by Proposition 5 of § 4; this excludes the possibility of representing metastable states. Moreover, Φ_ξ is not sequentially weakly lower semicontinuous; this causes difficulties in the study of the evolution problem, in which the phase equilibrium condition (5.1) is coupled with the heat diffusion equation, namely, the classical *Stefan problem* (cf., e.g., [7]). In the weak formulation of the latter problem, these difficulties are overcome by means of convex regularization, namely, by replacing Φ_ξ with Φ_ξ^{**}. This introduces as possible equilibrium states also those corresponding to $-1 < u < 1$; the zones where this occurs are interpreted as very fine mixtures of solid and liquid (so-called *mushy regions*), with liquid concentration equal to $(u+1)/2$.

Obviously the convex regularization of Φ_ξ does not introduce any relative minimum for the functional, so that the representation of metastable states is still excluded. Moreover, in several cases, conditions (5.1) are violated—either $u = 1$ and $\xi < 0$ (*supercooling*) or $u = -1$ and $\xi > 0$ (*superheating*) can occur. These effects are especially evident in *nucleation* phenomena, namely, in the formation of a new phase. For a sufficiently large single-phase system at a uniform temperature, once the duration of our observation has been prescribed, there are two critical values $a_1 = a_1(T) < 0$ and $a_2 = a_2(T) > 0$ such that any stationary configuration of the system fulfills the following conditions:

$$\begin{aligned} &\text{If } \xi \leqq -a_1, \text{ then } u \equiv -1 \text{ in } \Omega \text{ (“all solid”)}, \\ &\text{If } \xi \geqq a_2, \text{ then } u \equiv 1 \text{ in } \Omega \text{ (“all liquid”)}, \\ &\text{If } -a_1 < \xi < a_2, \text{ then both } u \equiv -1 \text{ in } \Omega \text{ and } u \equiv 1 \text{ in } \Omega \text{ are possible.} \end{aligned} \tag{5.3}$$

This phenomenon is due to the contribution given by the solid-liquid interface $\mathcal{S}$ to the global free enthalpy of the system [3, Chap. 3]. We represent this term by

$$\hat{\Lambda}(u) = \frac{\sigma}{2} \int_\Omega |\nabla u| + \frac{\sigma_L - \sigma_S}{2} \int_{\partial\Omega} u \, d\mathcal{H}_{N-1} + \frac{\sigma_L + \sigma_S}{2} \mathcal{H}_{n-1}(\partial\Omega); \tag{5.4}$$

here the constants σ, σ_L, and σ_S denote the surface tension coefficients for a solid-liquid interface and for contact between liquid and solid phases and an exterior substance, respectively; $\mathcal{H}_{N-1}$ denotes the $(N-1)$-dimensional Hausdorff measure. Note that $\hat{\Lambda}$ is equal to $(\sigma/2)V$ (cf. (2.7)) plus an affine term.

Thus our system is governed by the potential $\hat{\Psi}_\xi := \Phi_\xi + \hat{\Lambda}$. As we saw in § 4, $\hat{\Psi}_\xi$ admits not only an absolute minimum but also, for suitable ξ's, a relative one; moreover, each of these minima u is the characteristic function of some set $\Omega^+ \subset \Omega$ (i.e., $u = 1$ in Ω^+, $u = -1$ in $\Omega \backslash \Omega^+$).

The either absolute or relative minima of $\hat{\Psi}_\xi$ fulfill a contact angle condition between the interface $\mathcal{S}$ (i.e., the set of interior points of the relative boundary of Ω^+) and the boundary of Ω; this condition depends on the boundary terms of (5.4) (cf. [20, § 3]). These boundary terms also satisfy a weak form of the classical *Gibbs-Thomson law* (here for $\Omega \subset \mathbb{R}^3$)

$$\xi = -2\sigma\kappa \quad \text{on } \mathcal{S} \tag{5.5}$$

at the points of continuity of ξ on $\mathcal{S}$; here κ denotes the local mean curvature of $\mathcal{S}$ (assumed positive for a solid ball). Usually σ is so small that (5.2) is an acceptable

approximation of (5.5); however, the surface tension effect is more important where the interface is only potentially present, as in *nucleation* phenomena. In § 4 we interpreted all relative (nonabsolute) minima of the potential as metastable states of the system, namely, as states that persist only for some time and then decay because of thermodynamic fluctuations. Also note that for any $T>0$ there are relative minima of the potential representing states that cannot persist for a time T. For instance, if $\sigma_L=\sigma_S$, then for any $\xi\in L^\infty(\Omega)$ both states $u\equiv -1$ in Ω ("all solid") and $u\equiv 1$ in Ω ("all liquid") are either absolute or relative minima of $\hat{\Psi}_\xi$ cf. [20, § 3]. Here we want to give a criterion for selecting the metastable states that can persist for a prescribed time T. As we have already remarked, there exist two critical temperatures, $a_1(T)<0$ and $a_2(T)>0$, such that (5.3) holds; note that these values can be directly measured by experiments on a system at uniform temperature. Let us take $\tilde{\phi}$ such that

$$\begin{aligned}&\tilde{\phi}(-1)=\tilde{\phi}(1)=0,\quad \tilde{\phi}'(-1)=a_2,\quad \tilde{\phi}'(1)=-a_1,\\ &\tilde{\phi}\text{ is strictly concave in }[-1,1],\qquad \tilde{\phi}(u)=+\infty\quad\text{if } |u|>1;\end{aligned}\tag{5.6}$$

for instance, for $a_1=a_2$ $(=: a)$ we can take

$$\tilde{\phi}(u)=\begin{cases}\dfrac{a}{2}(1-u^2) & \text{if } |u|\leqq 1,\\ +\infty & \text{if } |u|>1\end{cases}\tag{5.7}$$

(cf. Fig. 2). We conjecture that for any temperature distribution $\xi:\Omega\to R$ the relative (nonabsolute) minima of the functional $\tilde{\Psi}_\xi:=\tilde{\Phi}_\xi+\Lambda$ correspond to the metastable states that persist for the time period T. Note that if ξ is uniform in space, then this criterion is obviously consistent with (5.3).

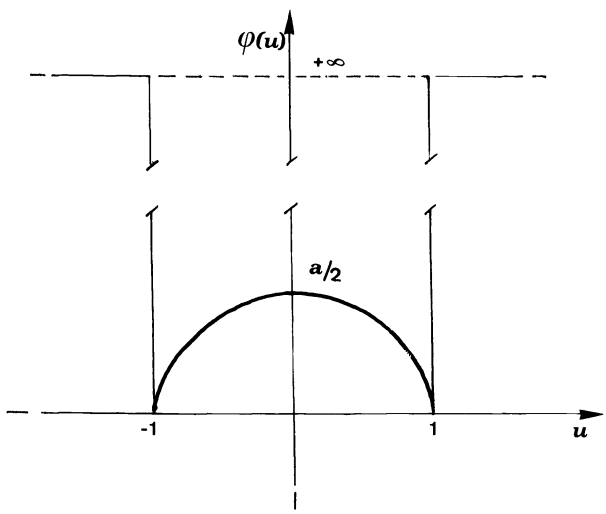


FIG. 2. *Free enthalpy density function of a solid-liquid substance at the relative temperature* $\theta=0$:

$$\tilde{\phi}(u):=\begin{cases}\dfrac{a}{2}(1-u^2) & \text{if } |u|\leqq 1,\\ +\infty & \text{if } |u|>1,\end{cases}$$

(a: *constant* >0). *Here* $u=-1$ ($u=1$, *respectively*) *corresponds to the solid phase* (*liquid phase, respectively*). *More generally, we can consider* $\tilde{\phi}$ *as in* (5.6). *For any* $\theta\in\mathbb{R}$, *setting* $\xi:=L\theta/2\tau_E$ (*with* L, τ_E: *constants greater than zero*), *the free-enthalpy density function is* $\tilde{\phi}_\xi(u):=\tilde{\phi}(u)-\xi u$ *for all* $u\in\mathbb{R}$.

Then it is not difficult to check that, for a system at a uniform temperature occupying a sufficiently large sample, the either absolute or relative minima of $\tilde{\Psi}_\xi$ fulfill (5.3). However, according to this model, in sufficiently small regions there is no limit to supercooling and superheating, a fact that seems physically plausible; we refer to [20, § 3] for details.

For any $\xi \in L^p(\Omega)$ with $p > N$ and for any absolute minimum u of $\tilde{\Psi}_\xi$, by (3.39) the corresponding "reduced interface" $\mathcal{S}^*$ is a $C^{1,(p-N)/2p}$ surface, by a classical result of De Giorgi [8]. The same result holds also for the relative minima of $\tilde{\Psi}_\xi$, since they are absolute minima of a suitably modified functional $\tilde{\Psi}_{\tilde{\xi}}$, by Theorem 4 of [20].

5.2. Selected mushy regions. So far we have excluded intermediate states between solid and liquid from our model. However, in some cases there is evidence of such states: *Mushy regions* can appear by internal heating of a solid; also, *clouds* and *fog* can be regarded as intermediate phases between liquid and vapor. These states correspond to very fine mixtures of the two pure phases; hence they are characterized by the concentration χ of the more energetic component, for instance, liquid in a solid-liquid system. As we set $u = -1$ in the solid and $u = 1$ in the liquid, any $\chi \in [0, 1]$ corresponds to $u = 2\chi - 1$. To represent intermediate phases in our model, it is sufficient to take a potential function ϕ that is locally convex on a suitable subset of $]-1, 1[$; thus we can select a precise concentration range for our stationary mushy regions.

Several choices are possible. If ϕ is as sketched in Fig. 3a, then two *stable* intermediate phases can be formed; they also appear as two layers between solid and liquid phases, a fact that does not seem to correspond to usual phenomena. For ϕ as in Fig. 3b, two *metastable* intermediate phases can be formed; they cannot appear at solid-liquid interfaces, which would seem nearer to the usual physical evidence.

A free enthalpy density ϕ similar to that of Fig. 3a, but with vertical asymptotes for $u = -1$ and $u = 1$, can represent systems of two partially miscible components; in this case $(u+1)/2$ represents the concentration of one of the two components [2].

5.3. Interfaces with infinite perimeter. The physical meaning of the phase-interaction functionals V and Λ_r $(0 < r < 1)$ can be roughly understood in the framework of the so-called *quasi-chemical approach* (see, e.g., [2, Chap. 2]). This theory is based on the assumption that there exist *bonds*, namely, forces, between pairs of atoms, and that these bonds have different intensities in the two phases. This accounts for the presence of a latent heat of phase transition. Bonds are also present between atoms of different phases; this corresponds to the presence of a phase-interaction term in the energy functional.

Usually, just the interaction between nearest neighbors is considered; then, as shown in [3, Chap. 2], the atoms at the interface between two phases contribute to the free enthalpy, which is proportional to the area of the interface itself. Thus, still setting $u = -1$ in the solid and $u = 1$ in the liquid, we can represent this quantity by

$$\frac{\sigma}{2} V(u) := \frac{\sigma}{2} \int_\Omega |\nabla u| \qquad (\leqq +\infty), \tag{5.8}$$

where $\sigma > 0$ is the surface tension coefficient. The contribution on the boundary of Ω (cf. (5.4)) can be similarly justified.

Let us now remove the restriction on bonds between nearest neighbors; then the zone of interaction between the two phases is not confined to their interface. If we denote by $4g(x, y)$ (greater than or equal to zero) the contribution due to two atoms of different phases sitting at two points $x, y \in \Omega$, then the global phase-interaction

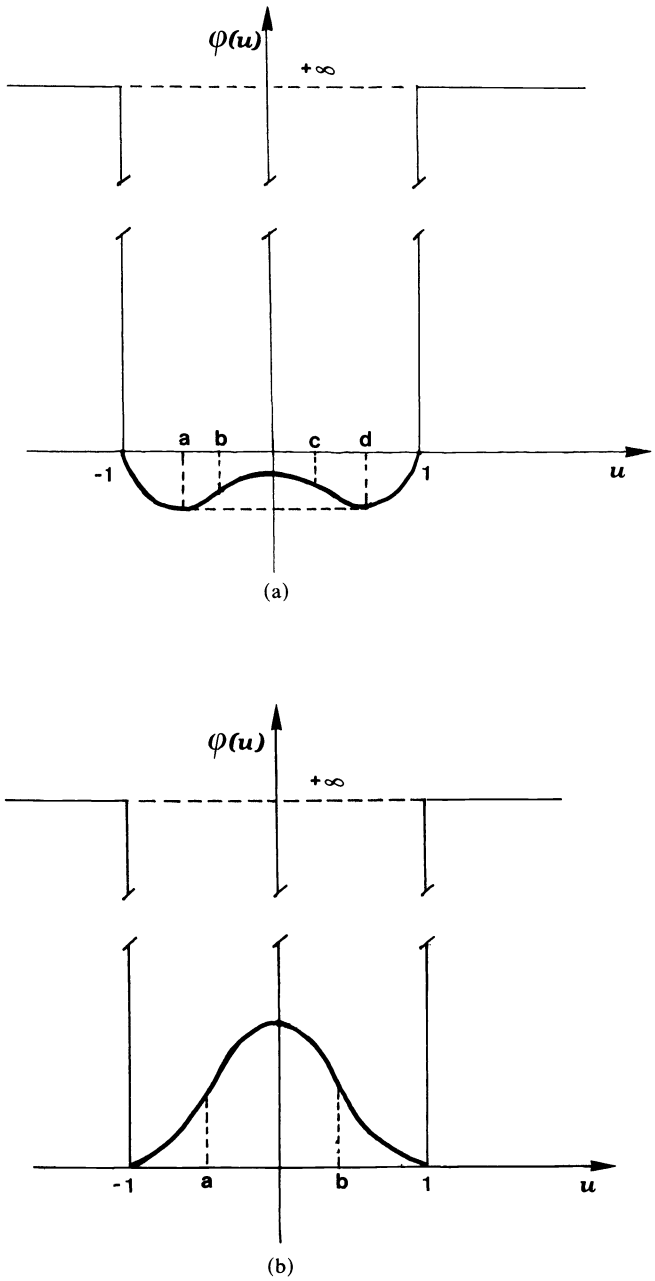


FIG. 3. *Examples of free-enthalpy density functions for systems with phase mixtures, corresponding to* $-1 < u < 1$.

(a) $\phi(a) = \phi(d) = \inf \phi$; $u = b$ *and* $u = c$ *are flexi.* $-1 \leqq u \leqq a$ *and* $d \leqq u \leqq 1$ *correspond to stable states;* $a < u < b$ *and* $c < u < d$ *to metastable ones.*

(b) $u = a$ *and* $u = b$ *are flexi.* $u = \pm 1$ *correspond to stable states;* $-1 < u < a$ *and* $b < u < 1$ *to metastable ones.*

The limit cases $b = c$ *in* (a) *and* $a = b$ *in* (b) *are admitted.*

contribution to the free enthalpy is given by

$$\Lambda_g(u) = \int\int_{\Omega^2} |u(x) - u(y)| g(x, y)\, dx\, dy \qquad (\leqq +\infty). \tag{5.9}$$

This model can be compared with that proposed by Rogers [17].

If the substance is homogeneous and isotropic, then $g(x, y) = \tilde{g}(|x - y|)$, with $\tilde{g}$ decreasing function $\mathbb{R}^+ \to \mathbb{R}^+$. If $\tilde{g}$ has large orders of infinity at zero and of infinitesimum at $+\infty$, then the functional Λ_g corresponds to a short-range interaction; consequently, the free-enthalpy contribution of the bulk of each phase is not much affected by the interaction with the other phase. In particular, if

$$\tilde{g}(h) = \sigma_r h^{-(3+r)} \quad \forall h \in \mathbb{R}^+ \tag{5.10}$$

(3 is the space dimension, σ_r: constant greater than zero), with $0 < r < 1$, then the corresponding functional Λ_r is a seminorm in the fractional Sobolev space $W^{r,1}(\Omega)$ (cf. [1] and § 2). Here it is not clear how the exterior boundary contribution to the free enthalpy might be represented, because the functions of $W^{r,1}(\Omega)$ $(0 < r < 1)$ do not admit traces on $\partial\Omega$.

$V(u)$ and $\Lambda_r(u)$ can be regarded as measures of the interface $\mathcal{S}$ between the two phases. For any $r \in]0, 1[$, if $V(u) < +\infty$, then $\Lambda_r(u) < +\infty$, but the converse does not hold. For instance, a *dendritic interface* can correspond to $\Lambda_r(u) < +\infty$ for some $r \in]0, 1[$; then $\mathcal{S}$ has infinite area, namely, $V(u) = +\infty$.

We point out two open questions:

(1) *Vectorial case.* Phase structures can also appear in systems characterized by a *vectorial* state variable. An example is given by the Landau-Lifshitz microscopic model of *ferromagnetism* [13, Chap. 5]. Here the state is characterized by the magnetization field, which has prescribed modulus and variable orientation; the phase structure corresponds to the splitting of the system into domains of approximately uniform magnetization. Some mathematical aspects of this phenomenon are considered in [19]. Here, what is the natural space interaction contribution to the free enthalpy?

(2) *Relation to the phase field model.* For $\phi := I_{\{-1,1\}}$, Modica and Mortola have shown that the absolute minima of $\Psi := \Phi + V$ can be characterized as the Γ-limit (in the sense of De Giorgi) of a family of more regular functionals [14]-[16]. Can a similar result be proved for V replaced by Λ_r $(0 < r < 1)$?

REFERENCES

[1] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] J. W. CAHN AND J. E. HILLIARD, *Free energy of a nonuniform system. I. Interfacial free energy*, J. Chem. Phys., 28 (1957), pp. 258-267.

[3] B. CHALMERS, *Principles of Solidification*, John Wiley, New York, 1964.

[4] I. EKELAND AND R. TEMAM, *Analyse convèxe et problèmes variationels*, Dunod, Gauthier-Villars, Paris, 1974.

[5] K. J. FALCONER, *The Geometry of Fractal Sets*, Cambridge University Press, Cambridge, U.K., 1985.

[6] W. H. FLEMING AND R. RISHEL, *An integral formula for total gradient variation*, Arch. Math., 11 (1960), pp. 218-222.

[7] A. FRIEDMAN, *The Stefan problem in several space variables*, Trans. Amer. Math. Soc., 133 (1968), pp. 51-87.

[8] E. GIUSTI, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Boston, 1984.

[9] M. E. GURTIN, ED., *Phase Transformations and Material Instabilities in Solids*, Academic Press, Orlando, FL, 1984.

[10] M. E. GURTIN, *On a theory of phase transitions with interfacial energy*, Arch. Rational Mech. Anal., 87 (1985), pp. 187-212.

[11] M. E. GURTIN, *On the two-phase Stefan problem with interfacial energy and entropy*, Arch. Rational Mech. Anal., 96 (1986), pp. 199-241.

[12] ———, *On phase transitions with bulk, interfacial, and boundary energy*, Arch. Rational Mech. Anal., 96 (1986), pp. 243-266.

[13] L. D. LANDAU AND E. M. LIFSHITZ, *Electrodynamics of Continuous Media*, Pergamon Press, Oxford, 1960.

[14] L. MODICA, *Gradient theory of phase transitions with boundary contact energy*, Ann. Inst. H. Poincaré Anal. non linéaire, 4 (1987), pp. 487-512.

[15] ———, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 48 (1987), pp. 123-142.

[16] L. MODICA AND S. MORTOLA, *Un esempio di Γ^--convergenza*, Bol. Un. Mat. Ital. B(6), 14 (1977), pp. 285-299.

[17] J. C. W. ROGERS, *The Stefan problem with surface tension*, in Free Boundary Problems: Theory and Applications, Vol. I, A. Fasano and M. Primicerio, eds., Pitman, Boston, 1983, pp. 263-274.

[18] A. VISINTIN, *Models for supercooling and superheating effects*, in Free Boundary Problems: Theory and Applications, Vols. III and IV, A. Bossavit, A. Damlamian, and M. Fremond, eds., Pitman, Boston, 1985, pp. 200-207.

[19] ———, *On Landau-Lifshitz equations for ferromagnetism*, Japan J. Appl. Math., 2 (1985), pp. 69-84.

[20] ———, *Surface tension effects in phase transitions*, in Material Instabilities in Continuum Mechanics and Related Mathematical Problems, J. M. Ball, ed., Clarendon Press, Oxford, 1988, pp. 505-537.

[21] ———, *Generalized coarea formula and fractal sets*, Japan J. Appl. Math., to appear.

[22] ———, *Surface tension effects in two phase systems*, in Free Boundary Problems: Theory and Applications, Vols. V and VI, K.-H. Hoffmann and J. Sprekels, eds., Longman, London, White Plains, NY, to appear.

[23] ———, *Pattern evolution*, Ann. Scuola Norm. Sup. Pisa, to appear.

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1305-1325, September 1990
© 1990 Society for Industrial and Applied Mathematics
012

NONADIABATIC PLANE LAMINAR FLAMES AND THEIR SINGULAR LIMITS*

VINCENT GIOVANGIGLI†

Abstract. A steady, premixed, nonadiabatic plane laminar flame with a one-step chemical mechanism is considered. By means of standard combustion approximations the model reduces to a two-point boundary value problem on the real line with an eigenvalue. Existence of a solution is achieved by first considering the problem in a bounded domain and by using the Leray-Schauder degree theory, and then by taking an infinite domain limit. The singular limit of high activation energy in the Arrhenius source term is then studied. Strong convergence to solutions of a limiting free boundary value problem determined by the discontinuity of the derivatives on the free boundary is proved. Further, correctors are specified in terms of the activation energy.

Key words. flame, nonadiabatic, wave

AMS(MOS) subject classifications. 34B15, 34B25, 34E20

1. Introduction. Traveling wave solutions to combustion models have recently been investigated by Berestycki, Nicolaenko, and Scheurer [1]. These authors have considered a system of two reaction diffusion equations modeling an adiabatic plane laminar flame with a nonunity Lewis number and a one-step chemical mechanism. Berestycki, Nicolaenko, and Scheurer have proved the existence of a solution and have studied the singular limit of high activation energy in the Arrhenius exponential term. They have proved strong convergence of the traveling wave to singular-limit free-boundary solutions with discontinuous derivatives. A related initial value problem has been investigated by Larrouturou [12] and Marion has considered a model with no ignition temperature [13]. In the framework of the constant density approximation [16] Berestycki and Larrouturou [2] and Berestycki and Niremberg [3] have also investigated a two-dimensional thermo-diffusive flame model.

In this paper, we present new results concerning nonadiabatic traveling waves and their singular limits. Nonadiabaticity is modeled by a heat loss term in the energy equation [16]. This situation fundamentally differs from the adiabatic case since multiple solutions are known to occur [4], [9], [16]. The crucial point in our analysis lies in the exchange of role between the reduced mass flux c of the wave, which is the natural eigenvalue of the problem, and the maximum heat loss rate parameter λ. A priori estimates for the reparameterized problem are derived and an upper bound c_{ad} for the reduced mass flux c is obtained. This upper bound corresponds to the adiabatic flame eigenvalue c_{ad}. Existence of a solution for $c \in (0, c_{\mathrm{ad}})$ is proved by first considering the problem in a bounded domain, which allows the reduction of the corresponding problem to a fixed-point formulation, and then by taking an infinite domain limit.

We then specify the dependence of the chemical Arrhenius source term on the inverse of the reduced activation energy ε, and we investigate the singular limit $\varepsilon \to 0$. In the adiabatic case, a one-term expansion has been rigorously established by Berestycki, Nicolaenko, and Scheurer for general Lewis numbers [1]. In the nonadiabatic case, formal asymptotic analyses obtained with matched expansions have been known for a long time [4], [9], [17]. An important result of these asymptotic investigations is that the curve $c \to \lambda$ is bell shaped. But these results are not rigorous from a

* Received by the editors July 13, 1988; accepted for publication (in revised form) November 6, 1989.
† Centre de Mathématiques Appliquées, Unité de Recherche Associée 756 du Centre National de la Recherche Scientifique, Ecole Polytechnique, 91128 Palaiseau Cedex, France.

mathematical point of view. In this paper, we first prove strong convergence of the nonadiabatic traveling wave to singular limit free-boundary solutions with discontinuous derivatives. We next investigate corrector terms, and we rigorously justify a two-term expansion in powers of ε. These corrector terms require the introduction of stretched variables. In particular, we fully justify for the first time the reactive internal layer analysis [4], [9], [17].

The governing equations are presented in § 2, a priori estimates are derived in § 3, and existence of a solution is obtained in § 4. The first-order asymptotic analysis is then given in § 5 and correctors are obtained in § 6.

2. Setting of the problem.

2.1. Governing equations. We consider a steady, premixed, nonadiabatic plane laminar flame propagating in a tube. We assume that the chemical mechanism is reduced to a single-step reaction of order $n>0$ governed by Arrhenius' law. Heat exchanges between the gases and their surroundings are modeled by a heat loss term in the energy equation [4], [9], [16]. Using standard combustion approximations, e.g., very subsonic speeds, constant specific heats, Lewis number unity, Fick's law for diffusion, etc., the problem can be reduced to the following governing equations:

$$-u''+cu'=f(u)v^n-\lambda g(u), \tag{2.1}$$

$$-v''+cv'=-f(u)v^n, \tag{2.2}$$

with the boundary conditions

$$u(-\infty)=0, \qquad v(-\infty)=1, \tag{2.3}$$

$$u(+\infty)=0, \qquad v'(+\infty)=0, \tag{2.4}$$

where u denotes the reduced temperature, v the reactant mass fraction, c the reduced mass flux, f the reduced source term, λ the reduced heat loss rate in the hot gases, and g the reduced heat loss rate function. In this model, the physical unknowns are u, v, and $c>0$ and the governing equations (2.1), (2.2) are conservation equations for energy and reactant mass fraction which include diffusion, convection, reaction, and heat loss terms. The boundary conditions (2.3), (2.4) mean that the incoming reactant is at the cold temperature and that heat losses freeze the chemical reaction behind the flame. We refer to [4], [9], [16] for more details on the physical derivation of the model.

The natural problem would be to find a nontrivial solution (u, v, c), with $(u, v)\neq(0, 1)$ and $c>0$. However, formal asymptotic theories indicate the presence of turning points with respect to λ and therefore of multiple solutions [4], [9], [16], [17]. These theories also indicate that the reduced mass flux c may be used to reparameterize the solution's manifold so that λ does not depend monotonically on c. As a consequence, it is more convenient to consider c as a parameter and λ as the eigenvalue and to seek solutions (u, v, λ) of (2.1)-(2.4) with $(u, v)\neq(0, 1)$ and $\lambda>0$. This reparameterization technique closely follows continuation methods used for computing turning points [5], [6], [11], [15]. Note that from a physical point of view these turning points with respect to λ correspond to flammability limits. More generally, most laminar flame extinction limits are turning points [5], [6], [16].

2.2. Assumptions. The source term is usually given by Arrhenius' law and will be specified in § 5. For the existence result we will only assume that

$$f \text{ is } C^1[0, 1] \quad \text{and} \quad f \text{ is nondecreasing,} \tag{2.5}$$

$$\exists\theta\in(0, 1) \quad f=0 \text{ on } [0, \theta] \quad \text{and} \quad f>0 \text{ on } (\theta, 1]. \tag{2.6}$$

Here θ is an ignition temperature related to the cold boundary difficulty [1], [7], [8], [13]. Although the solutions of (2.1)-(2.4) will be such that $0 \leqq u \leqq 1$ we extend f to be defined on $\mathbb{R}$ by setting

$$\forall u \geqq 1 \quad f(u) = f(1) = m, \qquad \forall u \leqq 0 \quad f(u) = 0, \tag{2.7}$$

where m denotes the maximum of f on $\mathbb{R}$.

The heat loss function g satisfies the following hypotheses:

$$g \text{ is } C^1[0,1], \quad g(0) = 0, \quad g(1) = 1, \tag{2.8}$$

$$\exists \alpha, \beta \quad 0 < \alpha \leqq g' \leqq \beta. \tag{2.9}$$

The reduced heat loss term $\lambda g(u)$ is usually a polynomial in u whose coefficients depend on the geometry of the tube where the deflagration takes place [9], [16]. We extend g to be defined on $\mathbb{R}$ by setting

$$\forall u \geqq 1 \quad g(u) = g(1) + g'(1)(u-1), \qquad \forall u \leqq 0 \quad g(u) = -g(-u). \tag{2.10}$$

Last, for $n > 0$ we substitute $|v|^{n-1}v$ for v^n, we define $\bar{n} = \min(1, n)$ so that $v \to v^n$ is $C^{\bar{n}}(\mathbb{R}_+)$, and we assume that

$$c > 0, \tag{2.11}$$

since in scaled variables c represents the reactant mass flux.

2.3. Reduction to a problem on $\mathbb{R}_+$. Let (u, v, λ) be a nontrivial solution of (2.1)-(2.4). Then, after a shift of the origin, if need be, u satisfies

$$u(0) = \theta, \qquad \forall x < 0 \quad u(x) < \theta, \tag{2.12}$$

where θ is the ignition temperature (2.6). Indeed from Lemma A of the Appendix we easily check that $u \leqq \theta$ would imply that $(u, v) = (0, 1)$. Now from (2.6) we get that

$$v(x) = 1 + (v(0) - 1)\exp(cx), \tag{2.13}$$

$$v'(0) = c(v(0) - 1), \tag{2.14}$$

and

$$-u'' + cu' + \lambda g(u) = 0, \tag{2.15}$$

$$u(-\infty) = 0, \qquad u(0) = \theta, \tag{2.16}$$

and the boundary value problem (2.15), (2.16) is the object of the following proposition.

PROPOSITION 2.1. *Assume that $\bar{\theta}, \bar{\lambda} \geqq 0$, and $\bar{c} > 0$ are arbitrarily given constants. Then the problem*

$$-u'' + \bar{c}u' + \bar{\lambda} g(u) = 0$$

$$u(-\infty) = 0, \qquad u(0) = \bar{\theta}$$

has a unique solution u which depends continuously in the $C^3(\mathbb{R}_-)$ topology on $(\bar{\theta}, \bar{\lambda}, \bar{c})$, and the following inequality holds for $\bar{\theta} \geqq 0$:

$$\forall x \leqq 0 \quad \bar{\theta} \exp\left(x \frac{\bar{c} + \sqrt{\bar{c}^2 + 4\bar{\lambda}\beta}}{2}\right) \leqq u(x) \leqq \bar{\theta} \exp\left(x \frac{\bar{c} + \sqrt{\bar{c}^2 + 4\bar{\lambda}\alpha}}{2}\right). \tag{2.17}$$

Furthermore, if $\psi(\bar{\theta}, \bar{\lambda}, \bar{c}) = u'(0)$ then ψ is C^1 in $(\bar{\theta}, \bar{\lambda}, \bar{c})$, ψ is odd in $\bar{\theta}$ and

$$\frac{\bar{c} + \sqrt{\bar{c}^2 + 4\bar{\lambda}\alpha}}{2} \leqq \frac{\partial \psi}{\partial \bar{\theta}} \leqq \frac{\bar{c} + \sqrt{\bar{c}^2 + 4\bar{\lambda}\beta}}{2}. \tag{2.18}$$

Remark. The case $\bar{\theta} \leqq 0$ will be needed in § 4 where boundary conditions of mixed type $w'(0) = \psi(w(0), \bar{\lambda}, \bar{c})$ will be considered for arbitrary functions w.

Sketch of the proof. We first prove the existence of a solution. We do not consider the trivial cases $\bar{\lambda} = 0$ or $\bar{\theta} = 0$ and since g is odd we may assume that $\bar{\theta} > 0$ and $\bar{\lambda} > 0$. For $\gamma \in \mathbb{R}$ let us denote by u_γ the solution of the initial value problem

$$-u'' + cu' + \bar{\lambda} g(u) = 0,$$

$$u(0) = \bar{\theta}, \qquad u'(0) = \gamma,$$

and let us consider the subsets Γ_+ and Γ_- of $\mathbb{R}$ defined by

$$\Gamma_+ = \{\gamma \in \mathbb{R} / \ \exists x < 0,\ u_\gamma(x) < 0\},$$

$$\Gamma_- = \{\gamma \in \mathbb{R} / \ \exists x < 0,\ u_\gamma(x) > \bar{\theta}\}.$$

We claim that Γ_+ and Γ_- are nonempty open intervals of $\mathbb{R}$. Indeed Γ_+ and Γ_- are open from the continuous dependence of u_γ on γ, and from Lemma B of the Appendix, if $\gamma \in \Gamma_+$ then $[\gamma, +\infty) \subset \Gamma_+$ and similarly $\gamma \in \Gamma_-$ implies that $(-\infty, \gamma] \subset \Gamma_-$. Furthermore, from Lemma B and B' and (2.9) we easily check that $(\bar{\theta} r_\beta, +\infty) \subset \Gamma_+$ and $(-\infty, \bar{\theta} r_\alpha) \subset \Gamma_-$ where $r_\alpha = (\bar{c} + \sqrt{\bar{c}^2 + 4\bar{\lambda}\alpha})/2$ and $r_\beta = (\bar{c} + \sqrt{\bar{c}^2 + 4\bar{\lambda}\beta})/2$. Now let $\gamma \in \Gamma_+$, choose $x_0 < 0$ such that $u_\gamma(x_0) < 0$ and $u'_\gamma(x_0) > 0$, and let w be the solution of the initial value problem on $(-\infty, x_0]$

$$-w'' + cw' + \bar{\lambda}\alpha w = 0,$$

$$w(x_0) = u(x_0),\ w'(x_0) = u'(x_0).$$

Since $\forall s \leqq 0$, $g(s) \leqq \alpha s$ we deduce from Lemma B that $u_\gamma \leqq w$ and hence $u_\gamma \to -\infty$ as $x \to -\infty$. The same type of argument proves that if $\gamma \in \Gamma_-$ then $u_\gamma \to +\infty$ as $x \to -\infty$ and therefore Γ_- and Γ_+ are disjoint. As a consequence, there exists γ such that $0 \leqq u \leqq \bar{\theta}$. Denoting by u this solution we easily deduce from Lemma C of the Appendix that $\forall x \leqq 0$, $u'(x) > 0$. Hence $u(-\infty)$ exists so that $u'' - cu'$ has a finite limit when $x \to -\infty$ and this limit is necessarily zero; otherwise $u(-\infty) = \pm\infty$. Thus $\bar{\lambda} g(u(-\infty)) = 0$ and hence $u(-\infty) = 0$ since $\bar{\lambda} > 0$.

Now the uniqueness of the solution and (2.17) are consequences of Lemma A and the continuous dependence on $(\bar{\theta}, \bar{\lambda}, \bar{c})$ is straightforward. Furthermore, we can prove that $\psi = u'(0)$ is C^1 in $(\bar{\theta}, \bar{\lambda}, \bar{c})$ and that if Ψ is the unique solution of the boundary value problem

$$-\Psi'' + \bar{c}\Psi' + \bar{\lambda} g'(u)\Psi = 0,$$

$$\Psi(-\infty) = 0, \qquad \Psi(0) = 1,$$

then $\partial\psi/\partial\bar{\theta} = \Psi'(0)$ and inequalities (2.18) are easily obtained.

Now combining (2.13), (2.14), and Proposition 2.1 we get the following proposition.

PROPOSITION 2.2. *Every nontrivial solution* $(u, v) \neq (0, 1)$ *and* $\lambda \geqq 0$ *of* (2.1)-(2.4), *after a shift of the origin, is a solution of* (2.1), (2.2), (2.4), *and*

$$u(0) = \theta, \quad u'(0) = \psi(u(0), \lambda, c), \quad v'(0) = c(v(0) - 1), \tag{2.19}$$

and conversely every solution (u, v) *and* $\lambda \geqq 0$ *of* (2.1), (2.2), (2.4), (2.19) *can be extended to* $\mathbb{R}$ *by means of Proposition* 2.1 *and* (2.13) *in such a way that it is a nontrivial solution of* (2.1)-(2.4).

In the following we will denote $\mathcal{P}$ the problem of finding a solution (u, v, λ) to (2.1), (2.2), (2.4), (2.19) with $\lambda \geqq 0$.

3. A priori estimates.

3.1. A priori estimates for (u, v, λ). In the following statements, we derive strong estimates for solutions (u, v, λ) of $\mathcal{P}$, that is to say for solutions (u, v, λ) of (2.1), (2.2), (2.4), (2.19) such that $\lambda \geqq 0$.

PROPOSITION 3.1. *u and v satisfy*

$$0 \leqq v < 1, \qquad v < u + v < 1, \tag{3.1}$$

$$-c \leqq v' \leqq 0, \qquad -c \leqq u' + v' \leqq 0. \tag{3.2}$$

and λ is positive.

PROPOSITION 3.2. *There exists $x_\infty > 0$ such that*

$$\forall x \in [0, x_\infty) \qquad v'(x) < 0, \tag{3.3}$$

$$\forall x \in [x_\infty, +\infty) \qquad v'(x) = 0. \tag{3.4}$$

Furthermore either $u(x_\infty) = \theta$, $u < \theta$ on $(x_\infty, +\infty)$, and $0 < v(x_\infty) < 1$ or $v(x_\infty) = 0$. The latter case may only occur in the case $0 < n < 1$.

Proof. The proof of Propositions 3.1 and 3.2 is lengthy and technical but simple.

First note that $0 < v(0)$. Arguing by contradiction, we know from (2.19) that $v'(0) = c(v(0) - 1)$ so that $v(0) \leqq 0$ implies $v'(0) < 0$. Now letting $w(x) = v(x) - (v(0) + v'(0)x)$ we have $w(0) = 0$, $w'(0) = 0$ and $w''(0) = cv'(0) + f(u(0))v^n(0) < 0$. Thus the set $E = \{x > 0 / \forall t \in [0, x] w(t) \leqq 0,\ w'(t) \leqq 0\}$ is not empty. Letting now $z = \sup E$ we must have $z = +\infty$; otherwise $w(z) \leqq 0$, $w'(z) \leqq 0$ and $w''(z) = cv'(z) + f(u(z))v^n(z) < 0$ since $v'(z) = v'(0) + w'(z) < 0$ and $v(z) = v(0) + v'(0)z + w(z) < 0$ and w'' is negative in a neighborhood of z, which contradicts the definition of z. Thus we deduce that $v' \leqq v'(0)$ on $\mathbb{R}_+$, contradicting (2.4).

Now if $n \geqq 1$ then $v > 0$, whereas if $0 < n < 1$ then $v \geqq 0$. Arguing by contradiction, we assume that there exists x_0 such that $v(x_0) < 0$. Then since $0 < v(0)$, there exists x_1 such that $v(x_1) < 0$ and $v'(x_1) < 0$, and proceeding as above yields that $v' \leqq v'(x_1)$ on $[x_1, +\infty)$, contradicting (2.4). Therefore $v \geqq 0$, and in the case $n \geqq 1$ if $v(x_0) = 0$ and $v'(x_0) < 0$ we again get that $v' \leqq v'(x_0)$; whereas if $v(x_0) = 0$ and $v'(x_0) = 0$ then $v = 0$ from the uniqueness of the solution to the Cauchy-Lipschitz problem, contradicting (2.19).

We now claim that there exists $x_\infty > 0$ such that $v' < 0$ on $[0, x_\infty)$ and $v' = 0$ on $[x_\infty, +\infty)$. Indeed $v \geqq 0$ so that we have

$$v'' - cv' = f(u)v^n \geqq 0,$$

and from Lemma C of the Appendix, i.e., the strong maximum principle, applied on $[x, y]$ we deduce that either $v = \text{Cte}$ and $v'(x) = 0$ or $v(x) = \max_{[x,y]} v$ and $v'(x) < 0$, or $v(y) = \max_{[x,y]} v$ and $v'(y) > 0$. However, if $v(y) = \max_{[x,y]} v \geqq 0$ and $v'(y) > 0$ then letting $w(x) = v(x) - (v(y) + v'(y)(x - y))$ we have $w(y) = 0$, $w'(y) = 0$, and we may easily check that w'' remains positive on $[y, +\infty)$. Thus we have $v' \geqq v'(y)$ on $[y, +\infty)$, contradicting (2.4). Hence we have $v' \leqq 0$ and if $v'(x) = 0$ then for all $y \geqq x$, $v'(y) = 0$. Moreover, $v'(0) = 0$ implies $v' = 0$ on $\mathbb{R}$, and hence $v = 1$ and $u \leqq \theta$, which contradicts $u'(0) = \psi(\theta, \lambda, c) \geqq c\theta$. Therefore $v'(0) < 0$ and the set

$$E = \{x > 0 / \forall t \in [0, x) v'(t) < 0\}$$

is not empty. Furthermore E is bounded since there exists x_0 such that $u(x) \leqq \theta$ for $x \geqq x_0$ and hence such that $v'' - cv' = 0$ on $[x_0, +\infty)$, which implies that $v' = 0$ on $[x_0, +\infty)$ since $v'(+\infty) = 0$. Now letting $x_\infty = \sup E$, we obtain (3.3), (3.4). Moreover, since $v'(0) = c(v(0) - 1)$ and $v'(0) < 0$, we have $v \leqq v(0) < 1$ on $\mathbb{R}_+$.

Assuming then that $u(x_\infty) < \theta$, we deduce that $v' = 0$ in a neighborhood of x_∞ since $v'' - cv' = 0$ and $v'(x_\infty) = 0$, contradicting the definition of x_∞. On the other hand, if $u(x_\infty) > \theta$ then $f(u(x_\infty))v(x_\infty)^n = 0$ implies that $v(x_\infty) = 0$, which may only occur when $0 < n < 1$ from the uniqueness of the solution to the Cauchy-Lipschitz problem. Last, when $u(x_\infty) = \theta$ and $v(x_\infty) > 0$ we have $f(u(x)) = 0$ for $x \geqq x_\infty$ and hence $u(x) \leqq \theta$ and it is straightforward to prove that $u < \theta$ on $(x_\infty, +\infty)$ using the maximum principle. Hence, we have shown that either $u(x_\infty) = \theta$, for all $x > x_\infty$, $u(x) < \theta$, and $v(x_\infty) \in (0, 1)$, or $v(x_\infty) = 0$. The latter case may only occur when $0 < n < 1$.

We now claim that $u > 0$ on $\mathbb{R}_+$. Arguing by contradiction, we assume that there exists x_0 such that $u(x_0) = 0$ and $u > 0$ on $[0, x_0)$, so that $u'(x_0) \leqq 0$. But since $-u'' + cu' + \lambda g(u) = 0$ in a neighborhood of x_0, $u'(x_0) = 0$ implies that $u = 0$, contradicting $u(0) = \theta$. Hence $u'(x_0) < 0$ and proceeding as for v yields that $u(x) \leqq u(x_0) + u'(x_0)(x - x_0)$ on $[x_0, +\infty)$, contradicting $u(+\infty) = 0$.

Now using (2.1), (2.2), we get that

$$(u+v)'' - c(u+v)' = \lambda g(u) \geqq 0,$$

and from the maximum principle applied on $[x, y]$ with y large enough, we deduce that $(u+v)'(x) < 0$ since $u(x) > u(+\infty)$ and v is nonincreasing. Furthermore, from (2.19) we have

$$(u+v)'(0) = \psi(u(0), \lambda, c) + c(v(0) - 1) \geqq c(u(0) + v(0) - 1)$$

since $\psi(u(0), \lambda, c) \geqq cu(0)$. Therefore $u(0) + v(0) < 1$ since $u'(0) + v'(0) < 0$, and we have obtained that $u + v < 1$ and $u' + v' < 0$ on $\mathbb{R}_+$. Furthermore, u and v are such that $-c(1-v) \leqq v'$ and $-c(1-u-v) \leqq u' + v'$ on $\mathbb{R}_+$, as we can show by integrating the inequalities $v'' - cv' \geqq 0$ and $(u+v)'' - c(u+v)' \geqq 0$ from zero to x.

Finally, assuming that $\lambda = 0$, we have $-(u+v)'' + c(u+v)' = 0$, which implies that $u' + v' = \gamma \exp(cx)$, where γ is a constant. Then if $\gamma \neq 0$ we deduce with (2.4) that $u'(x) \to \pm\infty$ as $x \to +\infty$, implying that $u(x) \to \pm\infty$, which contradicts (2.4). We thus have $\gamma = 0$ so that $u + v$ is a constant. On the other hand, we get from (2.18) that $\psi(\theta, 0, c) = c\theta$ and using (2.19) we obtain $u'(0) + v'(0) = c(u(0) + v(0) - 1)$ so that finally $u + v = 1$ and $v(+\infty) = 1$ from (2.4), which contradicts $v(+\infty) < v(0) < 1$, and the proof is complete.

Remark. These propositions show how heat losses may freeze the chemical reaction, leading to a positive residual of unburnt reactant $v(x_\infty)$, unless the reaction has already been completed. This will occur in the case $0 < n < 1$ as will be seen later.

COROLLARY 3.3. *We have*

$$0 \leqq v(x_\infty) < v(0) < 1 - \theta, \tag{3.5}$$

$$0 < \lambda < \frac{c^2}{\alpha\theta^2}. \tag{3.6}$$

Proof. From Propositions 3.1 and 3.2 we deduce that $v(0) < 1 - \theta$ since $u(0) + v(0) < 1$ and $0 \leqq v(x_\infty) < v(0)$ since x_∞ is positive. On the other hand, we also deduce from $u'(0) + v'(0) \leqq 0$ that $\psi(\theta, \lambda, c) \leqq c$. Using (2.18), we get $\lambda < c^2/\alpha\theta^2$, and since $\lambda > 0$ from Proposition 3.1 we deduce (3.6).

3.2. An upper bound for the reduced mass flux c. In this section we first recall some results about adiabatic flames and then we prove that if the problem $\mathcal{P}$ has a solution, the reduced mass flux parameter c must be in some interval $(0, c_{ad})$. In the next sections we will show that this condition is sufficient.

When there are no heat exchanges between the hot gases and their surroundings the problem is called adiabatic and $\lambda = 0$. The corresponding equations are

$$-u'' + cu' = f(u)v^n, \tag{3.7}$$

$$-v'' + cv' = -f(u)v^n, \tag{3.8}$$

with the boundary conditions

$$u(-\infty) = 0, \qquad v(-\infty) = 1, \tag{3.9}$$

$$u(+\infty) = 1, \qquad v(+\infty) = 0. \tag{3.10}$$

The unknowns here are u, v, and $c > 0$. Since (3.7)-(3.10) imply that $u + v = 1$, we can prove as in Proposition 2.2 that (3.7)-(3.10) is equivalent to finding u defined on $\mathbb{R}_+$ and $c > 0$ such that

$$-u'' + cu' = F(u), \tag{3.11}$$

$$u(0) = \theta, \quad u'(0) = cu(0), \quad u(+\infty) = 1, \tag{3.12}$$

in the particular case $F(u) = f(u)(1-u)^n$. This adiabatic flame problem (3.11), (3.12) has been extensively studied [1], [7], [8], [10], [13], and we have the following theorem.

THEOREM 3.4. (*Johnson-Nachbar.*) *Assume that* $F \in C^0[\theta, 1]$, $F > 0$ *on* $(\theta, 1)$ *and* $F(1) = 0$. *Then there exists a unique solution* (u_{ad}, c_{ad}) of (3.11), (3.12), *the eigenvalue* c_{ad} *is positive, and the map sending* F *to* c_{ad} *is continuous for the* $C^0[\theta, 1]$ *topology and strictly increasing, i.e.,* $F_1 \leqq F_2$ *and* $F_1 \neq F_2$ *imply that* $0 < c_{ad}(F_1) < c_{ad}(F_2)$ *with obvious notation.*

We can now state the main result of this section.

PROPOSITION 3.5. *If there exists a solution of* $\mathcal{P}$ *then*

$$0 < c < c_{ad}, \tag{3.13}$$

where c_{ad} *is the adiabatic mass flux in the particular case* $F(u) = f(u)(1-u)^n$.

Proof. We use a phase plane argument. From Proposition 3.1 we know that $u < 1 - v$ and hence $f(u) \leqq f(1-v)$ so that

$$-v'' + cv' + f(1-v)v^n \geqq 0. \tag{3.14}$$

Let us introduce $k = (1/c)v' \circ v^{-1}$ from $[v(x_\infty), v(0)]$ to $\mathbb{R}_-$. From Proposition 3.2 and (3.14) we get that

$$\begin{aligned} &k \in C^0[v(x_\infty), v(0)] \cap C^1(v(x_\infty), v(0)], \\ &-k\frac{dk}{dy} + k + \frac{1}{c^2} f(1-y)y^n \geqq 0 \quad \text{on } (v(x_\infty), v(0)], \\ &k < 0 \quad \text{on } (v(x_\infty), v(0)], \quad k(v(x_\infty)) = 0, \quad k(v(0)) = v(0) - 1. \end{aligned} \tag{3.15}$$

Let us also introduce the solution (u_{ad}, c_{ad}) of (3.11), (3.12) with $F(u) = f(u)(1-u)^n$ and let us set $v_{ad} = 1 - u_{ad}$. Now introducing $k_{ad} = (1/c_{ad})v'_{ad} \circ v_{ad}^{-1}$ from $[0, 1-\theta]$ to $\mathbb{R}_-$ we can easily prove that

$$\begin{aligned} &k_{ad} \in C^0[0, 1-\theta] \cap C^1(0, 1-\theta], \\ &-k_{ad}\frac{dk_{ad}}{dy} + k_{ad} + \frac{1}{c_{ad}^2} f(1-y)y^n = 0 \quad \text{on } (0, 1-\theta], \\ &k_{ad} < 0 \quad \text{on } (0, 1-\theta], \quad k_{ad}(0) = 0, \quad k_{ad}(1-\theta) = -\theta. \end{aligned} \tag{3.16}$$

Now since $(d/dy)(k_{ad} - y) < 0$ and $0 \leqq v(x_\infty) < v(0) < 1 - \theta$ from (3.5) we get, by integrating from $v(0)$ to $1 - \theta$, that

$$k(v(0)) < k_{ad}(v(0)) < 0. \tag{3.17}$$

On the other hand, for $y \in (v(x_\infty), v(0)]$ we have from (3.15), (3.16)

$$\frac{d}{dy}\{(k^2 - k_{\mathrm{ad}}^2)G(y)\} \leqq 2\left(\frac{1}{c^2} - \frac{1}{c_{\mathrm{ad}}^2}\right) f(y) y^n G(y), \tag{3.18}$$

where

$$G(y) = \exp\left(\int_y^{v(0)} \frac{2}{k(s) + k_{\mathrm{ad}}(s)}\, ds\right),$$

and integrating (3.18) from $v(x_\infty)$ to $v(0)$ yields

$$k^2(v(0)) - k_{\mathrm{ad}}^2(v(0)) + k_{\mathrm{ad}}^2(v(x_\infty))G(v(x_\infty)) \leqq 2\left(\frac{1}{c^2} - \frac{1}{c_{\mathrm{ad}}^2}\right) \int_{v(x_\infty)}^{v(0)} f(y) y^n G(y)\, dy. \tag{3.19}$$

Now combining (3.17) and (3.19) we deduce that $c < c_{\mathrm{ad}}$, and the proof is complete since $c > 0$ from assumption (2.11).

4. Existence of a solution.

4.1. Existence in a bounded domain. In this section we consider a problem similar to $\mathcal{P}$ but posed on a bounded domain $[0, a]$. This problem is equivalent to a fixed-point equation that we solve with the Leray-Schauder degree theory. In the next section we will pass to the limit $a \to +\infty$. Our procedure follows that of Berestycki, Nicolaenko, and Scheurer [1].

We consider the following problem $\mathcal{P}_a$ where $0 < a < +\infty$ is given: assuming that $0 < c < c_{\mathrm{ad}}$ as in Proposition 3.5, find a solution (u, v, λ) defined on $[0, a]$ with $\lambda \geqq 0$ satisfying (2.1), (2.2), (2.19), and

$$u(a) = 0, \qquad v(a) = 0. \tag{4.1}$$

Let us introduce the Banach space $X = C^1[0, a] \times C^1[0, a] \times \mathbb{R}$ equipped with the norm $\|(u, v, \lambda)\| = \max(\|u\|_{C^1[0,a]}, \|v\|_{C^1[0,a]}, |\lambda|)$ and consider for $\tau \in [0, 1]$ the mapping K_τ from $C^1[0, a] \times C^1[0, a] \times \mathbb{R}_+$ to X defined by

$$K_\tau(u, v, \lambda) = (U, V, \lambda + \theta - U(0)), \tag{4.2}$$

where U and V are solutions of

$$-U'' + c_\tau U' = \tau(f(u)v^n - \lambda g(u)) + (1 - \tau)(mV - \lambda\alpha U), \tag{4.3}$$

$$-V'' + c_\tau V' = -\tau f(u)v^n - (1 - \tau)mV, \tag{4.4}$$

with the boundary conditions

$$U'(0) = \tau\psi(u(0), \lambda, c_\tau) + (1 - \tau)c_\tau U(0), \qquad V'(0) = c_\tau(V(0) - 1), \tag{4.5}$$

$$U(a) = 0, \qquad V(a) = 0, \tag{4.6}$$

and where c_τ denotes the constant

$$c_\tau = c\frac{c_{\mathrm{ad}}(F_\tau)}{c_{\mathrm{ad}}}, \tag{4.7}$$

with $F_\tau(u) = \tau F(u) + (1 - \tau)m(1 - u)$ and $c_{\mathrm{ad}} = c_{\mathrm{ad}}(F)$. Note that solutions of $\mathcal{P}_a$ are fixed points of K_1 and conversely.

PROPOSITION 4.1. *The operator K_τ is well defined and the map $(\tau, u, v, \lambda) \to K_\tau(u, v, \lambda)$ is compact.*

Sketch of the proof. U and V can be calculated explicitly, and the compactness is straightforward.

Now we introduce the open bounded set Ω of X defined by

$$\Omega = \{(u, v, \lambda) \in X / \|u\|_{C^1[0,a]} < R, \|v\|_{C^1[0,a]} < R, 0 < \lambda < \Lambda\}, \tag{4.8}$$

where $R > 0$, $\Lambda > 0$ are constants and the following proposition shows that the degree $d(I - K_\tau, \Omega, 0)$ can be defined for suitable R and Λ provided a is large enough.

PROPOSITION 4.2. *There exist constants R and Λ such that for every c in $(0, c_{ad})$, there exists $a_c > 0$ with*

$$\forall a \geqq a_c, \quad \forall \tau \in [0, 1] \quad (I - K_\tau)(\partial\Omega) \not\ni 0. \tag{4.9}$$

Proposition 4.2 relies on Lemmas 4.3–4.5 in which we derive strong estimates for fixed points of K_τ.

LEMMA 4.3. *Let (u, v, λ) be a fixed point of K_τ, then*

$$0 \leqq v \leqq 1, \qquad v \leqq u + v \leqq 1, \tag{4.10}$$

$$-c_\tau \leqq v' \leqq 0, \qquad -c_\tau \leqq u' + v' \leqq 0. \tag{4.11}$$

Proof. The proof is essentially similar to that of Proposition 3.1.

First note that $0 < v(0)$. Indeed from (4.5) we know that $v'(0) = c_\tau(v(0) - 1)$ so that $v(0) \leqq 0$ implies that $v'(0) < 0$. Letting now $w(x) = v(x) - (v(0) + v'(0)x)$ we have $w(0) = 0$, $w'(0) = 0$, and $w''(0) = c_\tau v'(0) + \tau f(u(0))v^n(0) + (1 - \tau)mv(0) < 0$. Thus the set $E = \{x \in (0, a] / \forall t \in [0, x] w(t) \leqq 0,\ w'(t) \leqq 0\}$ is not empty. Now letting $z = \sup E$ we must have $z = a$; otherwise $w(z) \leqq 0$, $w'(z) \leqq 0$, and $w''(z) = c_\tau v'(z) + \tau f(u(z))v^n(z) + (1 - \tau)mv(z) < 0$; since $v'(z) = v'(0) + w'(z) < 0$ and $v(z) = v(0) + v'(0)z + w(z) < 0$ and w'' is negative in a neighborhood of z, which contradicts the definition of z. Thus we get $v(a) \leqq v(0) + v'(0)a < 0$, contradicting (4.6).

To prove that $v \geqq 0$, we argue by contradiction and we assume that there exists x_0 with $v(x_0) < 0$. Then, since $0 < v(0)$, there exists $x_1 \in (0, x_0)$ such that $v(x_1) < 0$ and $v'(x_1) < 0$. Proceeding as above yields that $v(x) \leqq v(x_1) + v'(x_1)(x - x_1)$ on $[x_1, a]$, contradicting (4.6).

We now claim that $v' \leqq 0$ on $[0, a]$. Indeed, $v \geqq 0$ so that we have

$$v'' - c_\tau v' = \tau f(u)v^n + (1 - \tau)mv \geqq 0$$

and from Lemma C of the Appendix, i.e., the strong maximum principle, applied on $[x, a]$ we deduce that $v = \text{Cte}$ and $v'(x) = 0$ or $v'(x) < 0$ since $v(x) \geqq v(a) = 0$. Thus $v' \leqq 0$ on $[0, a]$ and from (4.5) we obtain that $c(v(0) - 1) = v'(0) \leqq 0$ and $v \leqq 1$.

To prove that $u \geqq 0$ on $[0, a]$ we again argue by contradiction. Indeed, if there exists $x_0 \in (0, a)$ such that $u(x_0) < 0$, then there exists $x_1 \in (0, x_0)$ with $u(x_1) < 0$ and $u'(x_1) < 0$ since $u(0) = \theta > 0$. Now letting $w(x) = u(x) - (u(x_1) + u'(x_1)(x - x_1))$ we have $w(x_1) = 0$ and $w'(x_1) = 0$ and $w''(x_1) = c_\tau u'(x_1) - \tau f(u(x_1))v^n(x_1) + \tau\lambda g(u(x_1)) - (1 - \tau)mv(x_1) + (1 - \tau)\lambda\alpha u(x_1) < 0$, since $u'(x_1) < 0 / v(x_1) \geqq 0$, and $u(x_1) < 0$. Thus the set $E = \{x \in (x_1, a] / \forall t \in [x_1, x] w(t) \leqq 0,\ w'(t) \leqq 0\}$ is not empty. Now letting $z = \sup E$, we must have $z = a$, otherwise $w(z) \leqq 0$, $w'(z) \leqq 0$, and $w''(z) = c_\tau u'(z) - \tau f(u(z))v^n(z) + \tau\lambda g(u(z)) - (1 - \tau)mv(z) + (1 - \tau)\lambda\alpha u(z) < 0$, since $u'(z) = u'(x_1) + w'(z) < 0$, $v(z) \geqq 0$, and $u(z) = u(x_1) + u'(x_1)(z - x_1) + w(z) < 0$, and w'' is negative in a neighborhood of z, which contradicts the definition of z. Thus $u(x) \leqq u(x_1) + u'(x_1)(x - x_1)$ on $[x_1, a]$, contradicting $u(a) = 0$.

Moreover, using (4.3), (4.4) we get that

$$(u + v)'' - c_\tau(u + v)' = \tau\lambda g(u) + (1 - \tau)\lambda\alpha u \geqq 0.$$

From the maximum principle applied on $[x, a]$, we deduce that $(u+v)'(x) \leqq 0$, since $(u+v)(x) \geqq (u+v)(a) = 0$. On the other hand, from (4.5) we have

$$(u+v)'(0) = \tau\psi(u(0), \lambda, c_\tau) + (1-\tau)c_\tau u(0) + c_\tau(v(0)-1) \geqq c_\tau(u(0)+v(0)-1),$$

since $\psi(u(0), \lambda, c_\tau) \geqq c_\tau u(0)$. Therefore $u(0)+v(0) \leqq 1$, since $u'(0)+v'(0) \leqq 0$, and we have obtained that $u+v \leqq 1$ and $u'+v' \leqq 0$ on $[0, a]$. Finally u and v are such that $-c_\tau(1-v) \leqq v'$ and $-c_\tau(1-u-v) \leqq u'+v'$ on $[0, a]$ as we can show by integrating the inequalities $v''-c_\tau v' \geqq 0$ and $(u+v)''-c_\tau(u+v)' \geqq 0$ from zero to x, and the proof is complete.

In the following lemma we obtain an upper bound for the eigenvalue λ.

LEMMA 4.4. *Let (u, v, λ) be a fixed point of K_τ; then*

$$\lambda < \frac{m}{\alpha\theta}. \tag{4.12}$$

Proof. Arguing by contradiction we assume that $\lambda \geqq (m/\alpha\theta)$. In this situation we easily check that $u \geqq \theta$ implies that $\tau(f(u)v^n - \lambda g(u))$ and $(1-\tau)(mv - \lambda\alpha u)$ are both nonpositive. Now letting $w(x) = u(x) - (u(0) + u'(0)x)$, we have $w(0) = 0$, $w'(0) = 0$, and $w''(0) = c_\tau u'(0) - \tau(f(\theta)v^n(0) - \lambda g(\theta)) - (1-\tau)(mv(0) - \lambda\alpha\theta) > 0$, since $u'(0) > 0$, $f(\theta)v^n(0) - \lambda g(\theta) \leqq 0$, and $mv(0) - \lambda\alpha\theta \leqq 0$. Moreover, we may easily show, as in the proof of Lemma 4.3, that w'' remains positive on $[0, a]$. This implies that $u \geqq \theta$ on $[0, a]$, which contradicts $u(a) = 0$.

The next lemma shows that the eigenvalue λ cannot be zero if a is large enough.

LEMMA 4.5. *For every c in $(0, c_{\mathrm{ad}})$ there exists $a_c > 0$ such that*

$$\forall a \geqq a_c, \quad \forall \tau \in [0, 1] \quad (I - K_\tau)(u, v, \lambda) = 0 \Rightarrow \lambda > 0. \tag{4.13}$$

Proof. Arguing by contradiction, we assume that for every positive integer i there exist $a_i \geqq i$, $\tau_i \in [0, 1]$, and u_i, $v_i \in C^1[0, a_i]$ such that $(u_i, v_i, 0)$ is a fixed point of K_{τ_i}. From Lemma 4.3 and (4.3), (4.4) we easily deduce that for every compact $[0, a]$, u_i and v_i are uniformly bounded in $C^2[0, a]$ for $i \geqq a$. Now using the Hölder continuity of $v \to v^n$, we further get that u_i and v_i are uniformly bounded in $C^{2+\bar{n}}[0, a]$ for $i \geqq a$, where $\bar{n} = \min(1, n)$. By compactness, possibly taking a subsequence, we may therefore assume that

$$\tau_i \to \tau, \qquad (u_i, v_i) \to (u, v) \quad \text{in } C^2_{\mathrm{loc}}(\mathbb{R}_+).$$

Now $\psi(\theta, 0, c) = c\theta$ from (2.18) so that $u_i'(0) = c_{\tau_i}\theta$ and a simple calculation gives that $u_i(x) + v_i(x) = 1 - \exp(c_{\tau_i}(x - a_i))$ on $[0, a_i]$. Taking the limit $i \to +\infty$ therefore yields

$$-u'' + c_\tau u' = \tau f(u)v^n + (1-\tau)mv,$$

$$u(0) = \theta, \quad u'(0) = c_\tau\theta, \quad u + v = 1,$$

which implies that u is nondecreasing and $u(+\infty) = 1$. Hence $c_\tau = c_{\mathrm{ad}}(F_\tau)$, which contradicts (4.7).

Proof of Proposition 4.2. From Theorem 3.4, $c_{\mathrm{ad}}(F_\tau)$ depends continuously on τ, and we may let $R > \sup\{c_{\mathrm{ad}}(F_\tau)/0 \leqq \tau \leqq 1\} + 1$ and $\Lambda = m/\alpha\theta$ and Lemmas 4.3–4.5 obviously imply (4.9).

The value of $d(I - K_\tau, \Omega, 0)$ is now given by the following proposition.

PROPOSITION 4.6. *Under the same hypotheses as Proposition* 4.2 *and provided a is large enough we have*

$$\forall \tau \in [0, 1] \quad d(I - K_\tau, \Omega, 0) = -1. \tag{4.14}$$

Proof. From the homotopy invariance of the degree we know that $d(I-K_\tau, \Omega, 0) = d(I-K_0, \Omega, 0)$. But K_0 is a mapping depending only on λ which reads

$$K_0(u, v, \lambda) = (U_0, V_0, \lambda + \theta - U_0(0)).$$

Introducing the homotopy

$$H_\tau(u, v, \lambda) = (\tau U_0, \tau V_0, \lambda + \theta - U_0(0)),$$

we may easily check that the map $(\tau, u, v, \lambda) \to H_\tau(u, v, \lambda)$ is compact. Moreover, if (u, v, λ) is a fixed point of H_τ, then we have $u = \tau U_0$, $v = \tau V_0$, and $U_0(0) = \theta$. Thus (U_0, V_0, λ) is a fixed point of K_0 and for a large enough we have from Proposition 4.2 that $\|U_0\|_{C^1[0,a]} < R$, $\|V_0\|_{C^1[0,a]} < R$, and $0 < \lambda < \Lambda$. Hence $\|u\|_{C^1[0,a]} = \tau\|U_0\|_{C^1[0,a]} < R$ and $\|v\|_{C^1[0,a]} = \tau\|V_0\|_{C^1[0,a]} < R$. We have thus shown that

$$\forall a \geqq a_c, \quad \forall \tau \in [0, 1] \quad (I - H_\tau)(\partial\Omega) \not\ni 0,$$

so that $d(I-H_\tau, \Omega, 0)$ is well defined and constant. Hence $d(I-H_1, \Omega, 0) = d(I-H_0, \Omega, 0)$ and since $H_1 = K_0$ we have $d(I-K_1, \Omega, 0) = d(I-H_0, \Omega, 0)$. Now H_0 is a mapping which reads

$$H_0(u, v, \lambda) = (0, 0, \lambda + \theta - U_0(0)),$$

and from the multiplicative property of the degree we deduce that

$$d(I-K_1, \Omega, 0) = d(U_0(0) - \theta, (0, \Lambda), 0).$$

To compute this degree, note that if

$$\phi(\lambda) = U_0(0),$$

then a lengthy but straightforward calculation leads to

$$\phi(\lambda) = \begin{cases} -\dfrac{\psi(\lambda) - \psi(m/\alpha)}{\lambda - (m/\alpha)} & \text{if } \lambda \neq \dfrac{m}{\alpha}, \\ -\psi'\left(\dfrac{m}{\alpha}\right) & \text{if } \lambda = \dfrac{m}{\alpha}, \end{cases} \tag{4.15}$$

where

$$\psi(t) = \frac{2mc_0/\alpha}{c_0 + \sigma(\sqrt{c_0^2 + 4\alpha t})}, \qquad \sigma(s) = s\frac{1+e^{-as}}{1-e^{-as}}. \tag{4.16}$$

This implies that

$$\phi(\lambda) = -\int_0^1 \psi'\left(\lambda + \left(\frac{m}{\alpha} - \lambda\right)r\right) dr,$$

and thus

$$\phi'(\lambda) = -\int_0^1 (1-r)\psi''\left(\lambda + \left(\frac{m}{\alpha} - \lambda\right)r\right) dr.$$

Now from (4.16) we deduce that

$$\psi''(t) = 8mc_0\alpha \frac{(c_0 + \sigma(s))(\sigma'(s) - s\sigma''(s)) + 2s\sigma'^2(s)}{s^3(c_0 + \sigma(s))^3},$$

where $s=\sqrt{c_0^2+4\alpha t}\in[c_0,\sqrt{c_0^2+4\Lambda\alpha}]$, but since $(\sigma'(s)-s\sigma''(s))=1+O(a^2e^{-ac_0})$ on the interval $[c_0,\sqrt{c_0^2+4\Lambda\alpha}]$, we get that $\psi''>0$ on $[0,\Lambda]$ for a large enough, so that $d\phi/d\lambda<0$ on $[0,\Lambda]$ for a large enough. On the other hand, we also note from (4.16) that

$$\psi=\frac{2mc_0}{c_0+\sqrt{c_0^2+4\alpha t}}+O(e^{-ac_0}) \quad \text{on } [0,\Lambda], \tag{4.17}$$

and an explicit calculation also yields

$$c_{\mathrm{ad}}(F_0)=(1-\theta)\sqrt{m/\theta}. \tag{4.18}$$

Now from (4.17) $\phi(0)=1-2c_0/(c_0+\sqrt{c_0^2+4m})+O(e^{-ac_0})$ and combining (4.7) and (4.18) gives that $c_0<(1-\theta)\sqrt{m/\theta}$ and hence $1-2c_0/(c_0+\sqrt{c_0^2+4m})>\theta$ so that $\phi(0)>\theta$ for a large enough. Similarly at $\Lambda=m/\alpha\theta$ we have $\phi(\Lambda)<\psi(m)/(\Lambda\alpha-m)$ so that $\phi(\Lambda)<(\theta/(1-\theta))(2c_0/(c_0+\sqrt{c_0^2+4m}))+O(e^{-ac_0})$ but since $c_0<(1-\theta)\sqrt{m/\theta}$ we again deduce that $(\theta/(1-\theta))\,(2c_0/(c_0+\sqrt{c_0^2+4m}))<\theta$ and that $\phi(\Lambda)<\theta$ for a large enough. Therefore $d(\phi(\lambda)-\theta,(0,\Lambda),0)=-1$, and the proof is complete.

The main result of this section is thus a direct consequence of Proposition 4.6.

THEOREM 4.7. *For all $c\in(0,c_{\mathrm{ad}})$ there exists a_c such that for any $a\geqq a_c$ the problem $\mathscr{P}_a$ has a solution.*

4.2. Existence of a solution.

THEOREM 4.8. *For $c\in(0,c_{\mathrm{ad}})$ the problem $\mathscr{P}$ has a solution.*

Proof. From Theorem 4.7 we know that, provided i is large enough, there exists a solution (u_i,v_i,λ_i) of $\mathscr{P}_i$ on $[0,i]$. From Lemmas 4.3 and 4.4 we easily check that λ_i is bounded and that for every compact $[0,a]$, u_i and v_i are uniformly bounded in $C^{2+\bar{n}}[0,a]$ for $i\geqq a$, where $\bar{n}=\min(1,n)$. From compactness, eventually taking a subsequence, we may assume that

$$\lambda_i\to\lambda\geqq 0, \qquad (u_i,v_i)\to(u,v) \quad \text{in } C^2_{\mathrm{loc}}(\mathbb{R}_+),$$

and passing to the limit we get that (u,v,λ) is a solution of (2.1), (2.2), (2.19) and that v and $u+v$ are nonnegative and nonincreasing. Hence $u(+\infty)$ and $v(+\infty)$ exist, and by integrating (2.2), we get that v' has a limit as $x\to+\infty$, which can only be zero, so that $v'(+\infty)=0$. Similarly, $(u+v)''-c(u+v)'$ has a limit as $x\to+\infty$, which can only be zero, so that $\lambda g(u(+\infty))=0$. On the other hand, $\lambda=0$ yields $u+v=1$ and $c=c_{\mathrm{ad}}$ contradicting $c<c_{\mathrm{ad}}$. Therefore, $\lambda>0$ so that $u(+\infty)=0$, (2.4) holds, and the proof is complete.

Remark. The uniqueness of a solution, which implies the regularity of $c\to\lambda$, is an open problem. Nevertheless, in the limit case of high activation energy, the asymptotic limit of the solution will be seen to be unique. However, in general, we observe the following proposition.

PROPOSITION 4.9. *Denoting $(u,v,\lambda(c))$ any solution of $\mathscr{P}$ for $c\in(0,c_{\mathrm{ad}})$, we have*

$$\lim_{c\to 0}\lambda(c)=0, \qquad \lim_{c\to c_{\mathrm{ad}}}\lambda(c)=0.$$

This proposition is a consequence of Corollary 3.3 and Proposition 3.5.

5. Asymptotic analysis.

5.1. Setting the problem. In this part we specify the Arrhenius source term $f=f_\varepsilon$ as

$$f_\varepsilon(u)=\mathscr{F}\frac{1}{\varepsilon^{n+1}}\exp\left(\frac{u-1}{\varepsilon}\right)\chi(u), \tag{5.1}$$

where $\varepsilon > 0$ is a parameter, $\mathcal{F} > 0$ is a constant, and χ satisfies

$$\chi \text{ is } C^1(\mathbb{R}) \quad \chi \text{ nondecreasing,} \tag{5.2}$$

$$\exists \theta' \in (\theta, 1) \quad \chi = 0 \quad \text{on } (-\infty, \theta] \quad \text{and} \quad \chi = 1 \quad \text{on } [\theta', +\infty). \tag{5.3}$$

In scaled variables, ε represents the inverse of the activation energy of the chemical reaction. In practice the activation energy is large motivating the analysis $\varepsilon \to 0$ [1], [16]. In the case of adiabatic flames, we have the following result [1].

THEOREM 5.1. (*Berestycki, Nicolaenko, and Scheurer.*) *Denoting* $F_\varepsilon(u) = f_\varepsilon(u)(1-u)^n$ *we have*

$$\lim_{\varepsilon \to 0} c_{\text{ad}}(F_\varepsilon) = \sqrt{2\mathcal{F}\Gamma(n+1)}, \tag{5.4}$$

where Γ *is the Euler function.*

Simply denoting by c_{ad} this limit value, $c < c_{\text{ad}}$ implies that $c < c_{\text{ad}}(F_\varepsilon)$ for small enough, say $\varepsilon < \varepsilon_0$, and therefore implies the existence of a corresponding solution $(u_\varepsilon, v_\varepsilon, \lambda_\varepsilon)$ with $f = f_\varepsilon$. The goal of this section is to study the asymptotic behavior of $(u_\varepsilon, v_\varepsilon, \lambda_\varepsilon)$ as $\varepsilon \to 0$. Formal multiscaled asymptotic expansions of $(u_\varepsilon, v_\varepsilon, \lambda_\varepsilon)$ as $\varepsilon \to 0$ have been known for a long time [4], [9], [17]. In particular, an important result of these asymptotic investigations is that

$$\frac{\lambda_\varepsilon}{\varepsilon} \sim -\frac{c^2 \log (c^2/c_{\text{ad}}^2)}{\int_0^1 (g(u)/u)\, du + 1},$$

which implies that the curve $c \to \lambda$ is bell shaped [4], [9], [17]. However, such formal expansions are not rigorous from a mathematical point of view. In the adiabatic flame case and for general Lewis numbers a one-term expansion has been rigorously established by Berestycki, Nicolaenko, and Scheurer [1]. Marion has also studied a model with no ignition temperature [13]. In this paper, we rigorously justify a two-term expansion in powers of ε of $(u_\varepsilon, v_\varepsilon, \lambda_\varepsilon)$. We first proved strong convergence of the nonadiabatic traveling wave to singular limit free boundary solutions with discontinuous derivatives. We next investigate corrector tems which require introducing stretched variables. In particular, we completely describe the internal layer, and we give the asymptotic expansion of λ_ε.

5.2. The one-term expansion.

THEOREM 5.2. *For every* $a > 0$, $u_\varepsilon - u_0$ *and* $v_\varepsilon - v_0$ *converge to zero in* $H^1(-\infty, a)$ *as* $\varepsilon \to 0$, *where* u_0 *and* v_0 *are the unique solution of*

$$-u_0'' + cu_0' = c\delta_{x=\bar{x}}, \tag{5.5}$$

$$-v_0'' + cv_0' = -c\delta_{x=\bar{x}}, \tag{5.6}$$

where $\delta_{x=\bar{x}}$ *is the Dirac measure at the point* $\bar{x}$, *with the boundary conditions*

$$u_0(-\infty) = 0, \quad v_0(-\infty) = 1, \quad u_0(0) = \theta, \tag{5.7}$$

$$u_0(+\infty) = 1, \quad v_0(+\infty) = 0, \tag{5.8}$$

and λ_ε *converges to zero.*

Remark. We may check easily that

$$u_0(x) = \begin{cases} \exp(c(x-\bar{x})) & \text{if } x \leqq \bar{x}, \\ 1 & \text{if } x \geqq \bar{x}, \end{cases} \qquad v_0 = 1 - u_0, \tag{5.9}$$

where $\bar{x} = -(\log \theta / c)$ is uniquely determined by the condition $u_0(0) = \theta$.

The proof of Theorem 5.2 relies on Lemmas 5.3 and 5.4. In Lemma 5.3 we first show that λ_ε is $O(\varepsilon)$ as $\varepsilon \to 0$.

LEMMA 5.3. *There exists $L>0$ such that for ε small enough we have*

$$0<\lambda_\varepsilon<L\varepsilon. \tag{5.10}$$

Proof. From (2.1)-(2.4) we deduce easily that

$$u_\varepsilon+v_\varepsilon=1-\lambda_\varepsilon h_\varepsilon, \tag{5.11}$$

where the reduced enthalpy h_ε is such that

$$h_\varepsilon=\frac{1}{c}\left\{\int_{-\infty}^{x} g(u_\varepsilon(t))\,dt+\int_x^{+\infty} g(u_\varepsilon(t))\exp(c(x-t))\,dt\right\}, \tag{5.12}$$

$$h'_\varepsilon=\int_x^{+\infty} g(u_\varepsilon(t))\exp(c(x-t))\,dt, \tag{5.13}$$

so that

$$\forall x\geqq 0 \quad \frac{1}{c}\int_{-\infty}^{0} g(u_\varepsilon(t))\,dt\leqq h_\varepsilon(0)\leqq h_\varepsilon(x). \tag{5.14}$$

Now thanks to Proposition 2.1 and Corollary 3.3 we have

$$\forall x\leqq 0 \quad \theta\exp(c(1+\sqrt{1+4\beta/\theta^2\alpha})x)\leqq u_\varepsilon(x) \tag{5.15}$$

and combining (5.14) and (5.15), we deduce the existence of a constant μ independent of ε such that for ε small enough

$$\forall x\geqq 0 \quad 0<\mu\leqq h_\varepsilon(x). \tag{5.16}$$

On the other hand, multiplying (2.2) by v'_ε and integrating from zero to $+\infty$ yields

$$\frac{1}{2}v'^2_\varepsilon(0)+c\int_0^{+\infty} v'^2_\varepsilon(t)\,dt=-\int_0^{+\infty} f_\varepsilon(u_\varepsilon(t))v^n_\varepsilon(t)v'_\varepsilon(t)\,dt,$$

but from Corollary 3.3 we know that $v'^2_\varepsilon(0)\geqq c^2\theta^2$ and from (5.1) we have

$$0\leqq -f_\varepsilon(u_\varepsilon)v^n_\varepsilon v'_\varepsilon\leqq \mathscr{F}\exp\left(-\frac{\lambda_\varepsilon}{\varepsilon}h_\varepsilon\right)\exp\left(-\frac{v_\varepsilon}{\varepsilon}\right)\left(\frac{v_\varepsilon}{\varepsilon}\right)^n\left(-\frac{v'_\varepsilon}{\varepsilon}\right),$$

so that using (5.16) we get

$$\frac{1}{2}c^2\theta^2\leqq \mathscr{F}\exp\left(-\frac{\lambda_\varepsilon}{\varepsilon}\mu\right)\Gamma(n+1),$$

and the proof is complete.

In the next lemma, we introduce an interval $[0, x_\varepsilon]$ such that $f_\varepsilon(u_\varepsilon(t))=O(\varepsilon)$ for $t\in[0, x_\varepsilon]$. This interval is similar to the one introduced in the adiabatic case [1].

LEMMA 5.4. *For ε small enough there exists a unique x_ε in $(0,+\infty)$ such that*

$$u_\varepsilon(x_\varepsilon)=1+(n+2)\varepsilon\log(\varepsilon) \quad \forall x\in(-\infty, x_\varepsilon) \quad u_\varepsilon(x)<u_\varepsilon(x_\varepsilon), \tag{5.17}$$

and x_ε is bounded by positive constants a_1 and a_2 independent of ε

$$0<a_1<x_\varepsilon<a_2.$$

Sketch of the proof. Consider the set

$$E=\{x>0/\forall t\in[0, x],\ u(t)\leqq 1+(n+2)\varepsilon\log(\varepsilon)\}.$$

For ε small enough, E is not empty since $u(0)=\theta<1+(n+2)\varepsilon\log(\varepsilon)$. Moreover, if $x\in E$, then $u_\varepsilon(t)\leqq 1+(n+2)\varepsilon\log(\varepsilon)$ on $[0,x]$ so that $|f_\varepsilon(u_\varepsilon(t))|\leqq \mathscr{F}\varepsilon$. Now letting $w(x)=u_\varepsilon(x)-(u_\varepsilon(0)+u'_\varepsilon(0)x)$, we have $w(0)=0$, $w'(0)=0$, and $w''(0)=cu'_\varepsilon(0)+O(\varepsilon)>0$ for ε small enough, say, for $(\mathscr{F}+L)\varepsilon<c^2\theta$. Thus the set $\bar{E}=\{s\in(0,x]/\forall t\in[0,s],\ w(t)\geqq 0,\ w'(t)\geqq 0\}$ is not empty. Now if $z=\sup\bar{E}$, we must have $z=x$, otherwise $w(z)\geqq 0$, $w'(z)\geqq 0$, and $w''(z)=cu'_\varepsilon(z)+O(\varepsilon)>0$, since $u'_\varepsilon(z)=u'_\varepsilon(0)+w'(z)\geqq c\theta$ and $|f_\varepsilon(u_\varepsilon(z))|+|\lambda g(u_\varepsilon(z))|\leqq(\mathscr{F}+L)\varepsilon<c^2\theta$. Thus we have $c\theta\leqq u'_\varepsilon(t)$ on $[0,x]$ and by integration, $c\theta x\leqq u_\varepsilon(x)<1$, and E is bounded. We can now set $x_\varepsilon=\sup E$ and integrating the inequality $u'_\varepsilon\leqq c$ also yields that $1+(n+2)\varepsilon\log(\varepsilon)-\theta\leqq cx_\varepsilon$, and the proof is complete.

Proof of Theorem 5.2. Thanks to Lemma 5.3, $\lambda_\varepsilon\to 0$ as $\varepsilon\to 0$, and now the proof is essentially similar to the adiabatic case [1]. Denoting a_1 and a_2 positive constants bounding x_ε for ε small enough and $a>0$ with $0<a_1<a_2<a$, we get from Proposition 3.1 that u_ε and v_ε are uniformly bounded in $C^1[0,a]$. Moreover, it can easily be shown that h_ε defined in (5.12) is also uniformly bounded on $[0,a]$ so that

$$u_\varepsilon+v_\varepsilon=1+O(\varepsilon)\quad\text{on } [0,a].$$

Now from compactness we may extract a subsequence ε_i, $i\geqq 0$, decreasing to zero such that $(x_{\varepsilon_i},u_{\varepsilon_i},v_{\varepsilon_i})$ converges to (x_*,u_*,v_*) in $[a_1,a_2]\times C^0[0,a]\times C^0[0,a]$. Taking the limit $i\to+\infty$, we obtain that $0\leqq u_*\leqq 1$, $u_*(x_*)=1$, $u_*+v_*=1$, and that v_* is nonincreasing. Hence $u_*=1-v_*$ is nondecreasing and $u_*(x)=1$ if $x\geqq x_*$. On the other hand, if $x<x_*$ and if i is large enough, then $x<x_{\varepsilon_i}$ and from Lemma 5.4 and (5.1) we get that $f_\varepsilon(u_\varepsilon)=O(\varepsilon)$ and $f'_\varepsilon(u_\varepsilon)=O(1)$ on $[0,x]$. Therefore u_ε and v_ε are uniformly bounded in $C^{2+\bar{n}}[0,x]$. Consequently, u_{ε_i} and u_{ε_i} converge in $C^2[0,x]$, and we easily deduce that

$$-u''_*-cu'_*=0\quad\text{on } [0,x_*),$$
$$u_*(0)=\theta,\qquad u'_*(0)=\psi(\theta,0,c)=c\theta,$$

so that $(x_*,u_*,v_*)=(\bar{x},u_0,v_0)$, and since the limit is unique $(x_\varepsilon,u_\varepsilon,v_\varepsilon)$ converges to $(\bar{x},u_0,v_0)$ in $[a_1,a_2]\times C^0[0,a]\times C^0[0,a]$ and the convergence in $H^1(-\infty,a)$ is straightforward.

Remark. For $x<\bar{x}$, u_ε and v_ε converge to u_0 and v_0, respectively, in $C^2(-\infty,x]$. Note also that v_ε converges to v_0 in $C^0(\mathbb{R})$ since v_ε is nonincreasing and nonnegative.

The preceding theorem shows that although the heat loss term is an $O(\varepsilon)$ perturbation, the reduced mass flux c may take any value between zero and c_{ad}.

5.3. The asymptotic value of $\lambda_\varepsilon/\varepsilon$.

THEOREM 5.5. *The asymptotic value $\mathscr{L}$ of $\lambda_\varepsilon/\varepsilon$ is given by*

$$\lim_{\varepsilon\to 0}\frac{\lambda_\varepsilon}{\varepsilon}=\mathscr{L}=-\frac{c^2\log(c^2/c_{\text{ad}}^2)}{\int_0^1(g(u)/u)\,du+1}. \tag{5.18}$$

Remark. Theorem 5.5 shows that the asymptotic curve $c\to\lambda_\varepsilon$ is bell shaped, so that from an heuristic point of view, there is a turning point with respect to λ that corresponds to an extinction limit. More generally, most laminar flame extinction limits are turning points [5], [6], [16].

The proof of Theorem 5.5 relies on Lemmas 5.6 and 5.7. In Lemma 5.6 we determine the limit as $\varepsilon\to 0$ of a useful integral.

LEMMA 5.6. *Let x_1 and x_2 be such that $0\leqq x_1<\bar{x}<x_2$; then*

$$\lim_{\varepsilon\to 0}\int_{x_1}^{x_2}f_\varepsilon(u_\varepsilon(t))v_\varepsilon^n(t)v'_\varepsilon(t)\,dt=-\frac{c^2}{2}. \tag{5.19}$$

Proof. Let $\varphi \in C^2[0, x_2]$ such that $\mathrm{Supp}\,(\varphi) \in (0, x_2)$. Then we have

$$\int_0^{x_2} f_\varepsilon(u_\varepsilon(t)) v_\varepsilon^n(t)\varphi(t)\,dt = \int_0^{x_2} v_\varepsilon(t)(\varphi''(t) + c\varphi'(t))\,dt,$$

and from Theorem 5.2 we deduce that

$$\lim_{\varepsilon\to 0} \int_0^{x_2} f_\varepsilon(u_\varepsilon(t)) v_\varepsilon^n(t)\varphi(t)\,dt = \int_0^{x_2} v_0(t)(\varphi''(t) + c\varphi'(t))\,dt = c\varphi(\bar{x}). \tag{5.20}$$

On the other hand, assuming that $0 \leqq \varphi \leqq 1$ and that $\varphi = 1$ in a neighborhood of $\bar{x}$, we have

$$\int_0^{x_2} f_\varepsilon(u_\varepsilon(t)) v_\varepsilon^n(t)\varphi(t)\,dt \leqq \int_0^{x_2} f_\varepsilon(u_\varepsilon(t)) v_\varepsilon^n(t)\,dt = [v'_\varepsilon - cv_\varepsilon]_0^{x_2},$$

so that we have

$$cv_\varepsilon(x_2) + \int_0^{x_2} f_\varepsilon(u_\varepsilon(t)) v_\varepsilon^n(t)\varphi(t)\,dt - c \leqq v'_\varepsilon(x_2) \leqq 0. \tag{5.21}$$

Combining (5.20) and (5.21) we now get that

$$\lim_{\varepsilon\to 0} v'_\varepsilon(x_2) = 0,$$

and passing to the limit in the following identity yields the desired result

$$\frac{1}{2} v'^2_\varepsilon(x_2) - \frac{1}{2} v'^2_\varepsilon(x_1) - c\int_{x_1}^{x_2} v'^2_\varepsilon(t)\,dt = \int_{x_1}^{x_2} f_\varepsilon(u_\varepsilon(t)) v_\varepsilon^n(t) v'_\varepsilon(t)\,dt,$$

keeping in mind that v_ε converge in $H^1[0, x_2]$ and $C^2[0, x_1]$.

In the next lemma we investigate the asymptotic behavior of $v_\varepsilon(x)$ as $\varepsilon \to 0$ when $x > \bar{x}$.

LEMMA 5.7. *Let $x_2 > \bar{x}$ and let $e(n) = +\infty$ if $n \leqq 1$ and $e(n) = (n+1)/(n-1)$ if $n > 1$. Then*

$$\forall s \in [1, e(n)) \quad \lim_{\varepsilon\to 0} \frac{v_\varepsilon(x_2)}{\varepsilon^s} = 0. \tag{5.22}$$

Proof. Let $Z_\varepsilon = v_\varepsilon/\varepsilon$ and $x'_2 \in (\bar{x}, x_2)$. From the proof of Lemma 5.3, since h_ε is uniformly bounded on $[x'_2, x_2]$ and since $\chi(u_\varepsilon) = 1$ on $[x'_2, x_2]$ for ε small enough, we deduce that there exist constants κ_1 and κ_2 such that

$$-\int_0^{x'_2} f_\varepsilon(u_\varepsilon(t)) v_\varepsilon^n(t) v'_\varepsilon(t)\,dt \leqq \kappa_1 \int_{Z_\varepsilon(x'_2)}^{+\infty} e^{-s} s^n\,ds, \tag{5.23}$$

and

$$0 \leqq \int_{Z_\varepsilon(x_2)}^{Z_\varepsilon(x'_2)} e^{-s} s^n\,ds \leqq -\kappa_s \int_{x'_2}^{x_2} f_\varepsilon(u_\varepsilon(t)) v_\varepsilon^n(t) v'_\varepsilon(t)\,dt, \tag{5.24}$$

so that $Z_\varepsilon(x'_2)$ is bounded from Lemma 5.6 and $Z_\varepsilon(x'_2) - Z_\varepsilon(x_2)$ converges to zero. Furthermore, from (2.2) we may write that

$$Z''_\varepsilon - cZ'_\varepsilon = \phi \frac{Z_\varepsilon^n}{\varepsilon^k}, \tag{5.25}$$

where $k=2$ and where ϕ is bounded by positive constants $0<\gamma\leqq\phi\leqq\delta$. Arguing now by contradiction we assume that $Z_\varepsilon^n(x_2)/\varepsilon^2$ does not converge to zero. From (5.23), (5.24) we may then extract a subsequence ε_i, $i\geqq 0$, such that $Z_{\varepsilon_i}(x_2)$ converges to a constant Z_0 and such that $Z_{\varepsilon_i}^n(x_2)/\varepsilon_i^2\geqq\mu>0$ where μ denotes a positive constant. From Proposition 3.1, Z_{ε_i} is nonincreasing so that we have $Z_{\varepsilon_i}(x_2)\leqq Z_{\varepsilon_i}\leqq Z_{\varepsilon_i}(x_2')$ on $[x_2', x_2]$ and hence Z_{ε_i} converges uniformly on $[x_2', x_2]$ to Z_0. On the other hand, from Proposition 3.1, $Z_{\varepsilon_i}^n/\varepsilon_i^2$ is also nonincreasing so that $Z_{\varepsilon_i}''-cZ_{\varepsilon_i}'\geqq\gamma\mu>0$ on $[x_2', x_2]$. But it implies that $Z_{\varepsilon_i}\leqq X_{\varepsilon_i}$ where $X_{\varepsilon_i}''-cX_{\varepsilon_i}'=\gamma\mu>0$, $X_{\varepsilon_i}(x_2')=Z_{\varepsilon_i}(x_2')$, and $X_{\varepsilon_i}(x_2)=Z_{\varepsilon_i}(x_2)$. Taking now the limit $i\to+\infty$ we deduce that $Z_0\leqq X_0$ where $X_0(x_2')=X_0(x_2)=Z_0$ and $X_0''-cX_0'=\gamma\mu>0$, an obvious contradiction. Therefore $Z_\varepsilon/\varepsilon^{(2/n)}$ converges to zero. Defining now a new $Z_\varepsilon=V_\varepsilon/\varepsilon^{(1+2/n)}$ and a new $k=2/n$ we may again write an equation similar to (5.25) and proceeding in a similar way yields that $V_\varepsilon/\varepsilon^{(1+2/n+2/n^2)}$ converges to zero. An easy induction completes the proof since $e(n)=1+2\sum_{i\geqq 1}1/n^i$.

Proof of Theorem 5.5. We may easily check that h_ε converges to h_0 in $C^2_{\mathrm{loc}}(\mathbb{R})$ where

$$h_0=\frac{1}{c}\left\{\int_{-\infty}^{x} g(u_0(t))\,dt+\int_x^{+\infty} g(u_0(t))\exp\left(c(x-t)\right)dt\right\}, \tag{5.26}$$

so that from (5.10)

$$h_0(\bar{x})=\frac{1}{c^2}\left\{\int_0^1 (g(u)/u)\,du+1\right\}.$$

Now if $x_1<\bar{x}<x_2$ are in a neighborhood of $\bar{x}$, and if ε is small enough, we have

$$-\int_{x_1}^{x_2} f_\varepsilon(u_\varepsilon(t))v_\varepsilon^n(t)v_\varepsilon'(t)\,dt=\mathscr{F}\exp\left(-\frac{\lambda_\varepsilon}{\varepsilon}h_\varepsilon(y_\varepsilon)\right)\int_{V_\varepsilon(x_2)/\varepsilon}^{V_\varepsilon(x_1)/\varepsilon} e^{-s}s^n/ds,$$

where $x_1\leqq y_\varepsilon\leqq x_2$. But $(V_\varepsilon(x_1)/\varepsilon)\to+\infty$ from Theorem 5.2 and $(V_\varepsilon(x_2)/\varepsilon)\to 0$ from Lemma 5.7, and the result is straightforward using Lemma 5.6, since x_1 and x_2 can be chosen arbitrarily near $\bar{x}$.

6. Asymptotic correctors. In this section, we estimate the corrector terms $(u_\varepsilon-u_0)/\varepsilon$ and $(v_\varepsilon-v_0)/\varepsilon$. Estimating these corrector terms requires introducing stretched variables commonly used by physicists in the method of matched asymptotic expansions. These correctors terms behave very differently indeed in the preheat zone, say $x<\bar{x}$, in the flame zone, say $x\approx\bar{x}$, or in the burnt gases zone, say $x>\bar{x}$. However, the proofs are rather technical and thus will be omitted. We refer to [5] for more details.

6.1. A new origin. Before estimating $(u_\varepsilon-u_0)/\varepsilon$ and $(v_\varepsilon-v_0)/\varepsilon$ we have to eliminate artificial singularities at $\bar{x}$ for correctors due to a first choice of the origin defined by $u_\varepsilon(0)=\theta$. Therefore we introduce $\mathcal{O}_\varepsilon$ such that

$$v_\varepsilon(\mathcal{O}_\varepsilon)=1-\theta, \tag{6.1}$$

and we define

$$\tilde{u}_\varepsilon(x)=u_\varepsilon(x+\mathcal{O}_\varepsilon),\quad \tilde{v}_\varepsilon(x)=v_\varepsilon(x+\mathcal{O}_\varepsilon),\quad \tilde{h}_\varepsilon(x)=h_\varepsilon(x+\mathcal{O}_\varepsilon), \tag{6.2}$$

$$\tilde{x}_\varepsilon=x_\varepsilon+\mathcal{O}_\varepsilon,\qquad \tilde{x}_{\infty,\varepsilon}=x_{\infty,\varepsilon}+\mathcal{O}_\varepsilon. \tag{6.3}$$

Note that we have

$$\mathcal{O}_\varepsilon=-\frac{1}{c}\log\left(1+\frac{\lambda_\varepsilon h_\varepsilon(0)}{\theta}\right)\sim-\frac{\mathscr{L}h_0(\bar{x})}{c\theta}\varepsilon, \tag{6.4}$$

so that $(\tilde{u}_\varepsilon, \tilde{v}_\varepsilon, \tilde{x}_\varepsilon)$ converges to $(u_0, v_0, \bar{x})$ as $(u_\varepsilon, v_\varepsilon, x_\varepsilon)$, and furthermore

$$\tilde{u}'_\varepsilon(0) = \psi(\tilde{u}_\varepsilon(0), \lambda_\varepsilon, c), \quad \tilde{v}'_\varepsilon(0) = -c\theta, \quad \tilde{v}_\varepsilon(0) = 1-\theta. \tag{6.5}$$

In the following sections we estimate $(\tilde{u}_\varepsilon - u_0)/\varepsilon$ and $(\tilde{v}_\varepsilon - v_0)/\varepsilon$.

6.2. Correctors in the preheat zone. In Lemma 6.1, we describe the asymptotic behavior of $(\tilde{v}_\varepsilon - v_0)/\varepsilon$, from which we can deduce Corollary 6.2 by using (5.11).

LEMMA 6.1. *Let $k \geqq 1$ and $K > (n+k+1)/c$, then, as ε goes to zero, $(\tilde{v}_\varepsilon - v_0)/\varepsilon^k$ is bounded in $C^2(-\infty, \bar{x} + K\varepsilon \log \varepsilon]$.*

COROLLARY 6.2. *There exists a constant $K > 0$ such that $(\tilde{u}_\varepsilon - u_0)/\varepsilon + \mathscr{L}h_0$ and $(\tilde{v}_\varepsilon - v_0)/\varepsilon$ converge to zero in $C^2(-\infty, \bar{x} + K\varepsilon \log \varepsilon]$ where h_0 is as in (5.26).*

6.3. Correctors in the flame zone. Heuristically, near $\bar{x}$ the reaction term $f_\varepsilon(u_\varepsilon)v_\varepsilon$ is no longer negligible and there is an internal layer where the chemical reaction takes place. In this section we describe the structure of this internal layer by introducing the usual stretched variable ξ [4], [9]:

$$\xi = \bar{x} + \frac{x - \bar{x}}{\varepsilon}, \tag{6.6}$$

and by estimating $(\tilde{u}_\varepsilon - u_0)/\varepsilon$ and $(\tilde{v}_\varepsilon - v_0)/\varepsilon$ as functions of ξ.

THEOREM 6.3. *Let $w_\varepsilon(\xi) = \tilde{v}_\varepsilon(\bar{x} + \varepsilon(\xi - \bar{x}))/\varepsilon$ and $0 < \delta < \frac{1}{2}$. Then, as a function of ξ, w_ε converges to w_0 in $C^2[\bar{x} - 1/\varepsilon^\delta, +\infty)$ where w_0 is the unique solution of*

$$\frac{d^2 w_0}{d\xi^2} = \frac{c^2}{2\Gamma(n+1)} \exp(-w_0) w_0^n, \tag{6.7}$$

$$\lim_{\xi \to -\infty} (w_0(\xi) + c(\xi - \bar{x})) = 0, \qquad \lim_{\xi \to +\infty} w_0(\xi) = 0. \tag{6.8}$$

The proof is based on an integral formulation of the governing equation for w_ε, on a priori estimates, and on Lemma 6.1, which is used to match the flame zone with the preheat zone [5].

COROLLARY 6.4. *Let $\mathscr{W}_0 = -c(\xi - \bar{x})$ for $\xi \leqq \bar{x}$ and $\mathscr{W}_0 = 0$ for $\xi \geqq \bar{x}$. Then, as functions of ξ, $(\tilde{u}_\varepsilon - u_0)/\varepsilon + w_0 - \mathscr{W}_0 + \mathscr{L}h_0(\bar{x})$ and $(\tilde{v}_\varepsilon - v_0)/\varepsilon - w_0 + \mathscr{W}_0$ converge to zero in $C^0[\bar{x} - 1/\varepsilon^\delta, \bar{x} + 1/\varepsilon^{2\delta}]$.*

Remark. Letting $\xi = \bar{x}$ in Theorem 6.3 we obtain that $\lim_{\varepsilon \to 0} \tilde{v}_\varepsilon(\bar{x})/\varepsilon = w_0(\bar{x}) > 0$ and that $\lim_{\varepsilon \to 0} \tilde{v}'_\varepsilon(\bar{x}) = (dw_0/d\xi)(\bar{x}) \in (0, c)$. Moreover, from Corollaries 6.2 and 6.4 we deduce that the matching zone between the flame zone and the preheat zone is $x \in [\bar{x} - \varepsilon^{1-\delta}, \bar{x} + K\varepsilon \log \varepsilon]$.

6.4. Correctors in the burnt gases zone. In this section we estimate $(\tilde{u}_\varepsilon - u_0)/\varepsilon$ and $(\tilde{v}_\varepsilon - v_0)/\varepsilon$ behind the flame zone. As in §§ 6.2 and 6.3 we first study $\tilde{v}_\varepsilon/\varepsilon$ and then from (5.11) we deduce the behavior of $(\tilde{u}_\varepsilon - 1)/\varepsilon$. The convergence of $\tilde{v}_\varepsilon/\varepsilon$ to zero turns out to be very different depending on the order of the chemical reaction n as shown in the following proposition.

PROPOSITION 6.5.

- *If $0 < n < 1$, let $\tilde{x}_{\infty,\varepsilon}$ as in (3.3), (3.4), (6.3) and let $\xi_{\infty,\varepsilon} = \bar{x} + (\tilde{x}_{\infty,\varepsilon} - \bar{x})/\varepsilon$. Then*

$$\tilde{v}_\varepsilon(\tilde{x}_{\infty,\varepsilon}) = 0$$

and

$$\lim_{\varepsilon \to 0} \xi_{\infty,\varepsilon} = \bar{x} + \frac{1}{c} \int_0^{+\infty} \left(\frac{1}{\gamma(s)} - 1 \right) ds,$$

where γ is defined by

$$\gamma(s)=\left\{\int_0^s e^{-t}t^n\,dt/\Gamma(n+1)\right\}^{1/2}.$$

• *If $n=1$, for all $k\geqq 1$, there exists a constant K such that $\tilde{v}_\varepsilon/\varepsilon^k$ is bounded in $C^2[\bar{x}-K\varepsilon\log\varepsilon,+\infty)$. In particular, we have*

$$\lim_{\varepsilon\to 0}\tilde{v}_\varepsilon(\tilde{x}_{\infty,\varepsilon})/\varepsilon^k=0.$$

• *If $n>1$, for all k in $[1,(n+1)/(n-1)]$, there exists a constant K such that $\tilde{v}_\varepsilon/\varepsilon^k$ is bounded in $C^2[\bar{x}+K\varepsilon^{\{(n+1)-k(n-1)\}/2},+\infty)$. Moreover, if $\mathcal{V}_\varepsilon=\tilde{v}_\varepsilon/\varepsilon^{(n+1)/(n-1)}$, then for every $a>\bar{x}$, $\mathcal{V}_\varepsilon$ converges to $\mathcal{V}_0$ in $C^2[a,+\infty)$ where $\mathcal{V}_0$ is the unique solution of*

$$\mathcal{V}_0''-c\mathcal{V}_0'=\frac{c^2}{2\Gamma(n+1)}\exp\left(-\frac{\mathscr{L}}{c^2}(x-\bar{x})\right)\mathcal{V}_0^n,$$

with

$$\lim_{x\to\bar{x}_+}\mathcal{V}_0=+\infty,\qquad \lim_{x\to+\infty}\mathcal{V}_0'=0,$$

and in particular

$$\lim_{\varepsilon\to 0}\tilde{v}_\varepsilon(\tilde{x}_{\infty,\varepsilon})/\varepsilon^{(n+1)/(n-1)}=\mathcal{V}_0(+\infty)>0.$$

COROLLARY 6.6. *Let $[x_1,x_2]\subset(\bar{x},+\infty)$; then $(\tilde{u}_\varepsilon-u_0)/\varepsilon+\mathscr{L}h_0$ and $(\tilde{v}_\varepsilon-v_0)/\varepsilon$ converge to zero in $C^2[x_1,x_2]$.*

6.5. The cooling zone. Finally from Theorem 5.2 we know that $\tilde{u}_\varepsilon$ converges to 1 on every compact of $(\bar{x},+\infty)$ although $\tilde{u}_\varepsilon(+\infty)=0$. In this section we investigate how $\tilde{u}_\varepsilon$ decreases to zero in terms of the stretched variable ζ [9]

$$\zeta=\bar{x}+\varepsilon(x-\bar{x}). \tag{6.9}$$

THEOREM 6.7. *Let $\mathcal{U}_\varepsilon(\zeta)=\tilde{u}_\varepsilon(\bar{x}+(\zeta-\bar{x})/\varepsilon)$. Then, as a function of ζ, $\mathcal{U}_\varepsilon$ converges to $\mathcal{U}_0$ in $C^1[\bar{x}+\varepsilon,+\infty)$, where $\mathcal{U}_0$ is the unique solution of*

$$\frac{d\mathcal{U}_0}{d\zeta}=-\frac{\mathscr{L}}{c}g(\mathcal{U}_0), \tag{6.10}$$

$$\mathcal{U}_0(\bar{x})=1. \tag{6.11}$$

The proof again relies on a priori estimates and on an integral formulation of the governing equation for $\mathcal{U}_\varepsilon$ [5].

Remark. It is remarkable that the solution $(\tilde{u}_\varepsilon,\tilde{v}_\varepsilon,\lambda_\varepsilon)$ exhibits three length scales of interest, $O(1)$ in general, $O(\varepsilon)$ in the flame zone, and $O(1/\varepsilon)$ in the cooling zone.

Acknowledgment. I thank Professor H. Berestycki for interesting discussions concerning this material.

Appendix. In this Appendix, we state several lemmas used in the preceding sections.

LEMMA A. *Let I be a closed interval of $\mathbb{R}$, with endpoints a and b, $-\infty\leqq a<b\leqq+\infty$, and let H be a C^1 mapping from $I\times\mathbb{R}\times\mathbb{R}$ into $\mathbb{R}$. Assume that for all $x\in I$ and $z\in\mathbb{R}$ the mapping $y\to H(x,y,z)$ is nondecreasing. Let r and s in $C^2(I)$ and $M\geqq 0$ such that*

$$\forall x\in I\quad r''(x)\geqq H(x,r(x),r'(x)),\qquad r(a)\leqq s(a)+M,$$

$$\forall x\in I\quad s''(x)\leqq H(x,s(x),s'(x)),\qquad r(b)\leqq s(b)+M;$$

then

$$\forall x \in I \quad r(x) \leqq s(x) + M.$$

Proof. If is easy to check that the lemma is a consequence of the particular case $-\infty < a < b < +\infty$ and $M = 0$. For this case a proof can be found in Protter and Weinberger [14, pp. 18 and 48].

LEMMA B. *Assume that* $\bar{\theta}$, $\bar{\lambda} \geqq 0$, $\bar{c} > 0$, *and* γ *are arbitrarily given constants and that* $u \in C^2(\mathbb{R}_-)$ *satisfies*

$$-u'' + \bar{c}u' + \bar{\lambda} g(u) = 0,$$
$$u(0) = \bar{\theta}, \qquad u'(0) = \gamma.$$

Let ω_1 *and* ω_2 *be such that*

$$-\omega_1'' + \bar{c}\omega_1' + \bar{\lambda} g(\omega_1) \geqq 0, \qquad -\omega_2'' + \bar{c}\omega_2' + \bar{\lambda} g(\omega_2) \leqq 0,$$
$$\omega_1(0) \leqq \bar{\theta},\ \omega_1'(0) \geqq \gamma, \qquad \omega_2(0) \geqq \bar{\theta},\ \omega_2'(0) \leqq \gamma;$$

then

$$\forall x \leqq 0 \quad \omega_1(x) \leqq u(x) \leqq \omega_2(x), \qquad \omega_2'(x) \leqq u'(x) \leqq \omega_1'(x).$$

Proof. See Protter and Weinberger [14, pp. 26 and 49].

LEMMA B'. *Assume that* $\bar{\theta}$, $\bar{c} > 0$, *and* $\nu > 0$ *are arbitrarily given constants, and for* $\gamma \in \mathbb{R}$ *consider the initial value problem*

$$-u'' + \bar{c}u' + \nu u = 0,$$
$$u(0) = \bar{\theta}, \qquad u'(0) = \gamma,$$

Then, letting $r = (\bar{c} + \sqrt{\bar{c}^2 + 4\nu})/2$, *we have*

$$\text{if } \gamma = \bar{\theta} r \quad \lim_{x \to -\infty} u = 0,$$
$$\text{if } \gamma > \bar{\theta} r \quad \lim_{x \to -\infty} u = -\infty,$$
$$\text{if } \gamma < \bar{\theta} r \quad \lim_{x \to -\infty} u = +\infty.$$

Proof. The proof is obvious.

LEMMA C. *Let* $[a, b]$ *be an interval of* $\mathbb{R}$, u *in* $C^2[a, b]$, ϕ *in* $C^0[a, b]$, *and assume that*

$$\forall x \in [a, b] \quad u'' + \phi u' \geqq 0.$$

Then, if u *reaches its maximum in* (a, b), u *is constant, if* u *reaches its maximum at* $x = a$, u *is constant or* $u'(a) < 0$, *and if* u *reaches its maximum at* $x = b$, u *is constant or* $u'(b) > 0$.

Proof. See Protter and Weinberger [14, pp. 4 and 7].

REFERENCES

[1] H. BERESTYCKI, B. NICOLAENKO, AND B. SCHEURER, *Traveling wave solutions to combustion models and their singular limits*, SIAM J. Math. Anal., 16 (1985), pp. 1207–1242.

[2] H. BERESTYCKI AND B. LARROUTUROU, *A semi-linear elliptic equation in a strip arising in a two-dimensional flame propagation model*, in preparation.

[3] H. BERESTYCKI AND L. NIREMBERG, *Uniqueness and singular limit of the solution of a two-dimensional elliptic model for flame propagation*, in preparation.
[4] J. BUCKMASTER, *The quenching of a deflagration wave*, Combust. Flame, 26 (1976), pp. 151-162.
[5] V. GIOVANGIGLI, *Structure et extinction de flammes laminaires prémélangées*, Thèse d'etat, Université Paris 6, 1988.
[6] V. GIOVANGIGLI AND M. D. SMOOKE, *Adaptive continuation algorithms with applications to combustion problems*, Appl. Numer. Math., 5 (1989), pp. 305-331.
[7] W. E. JOHNSON, *On a first-order boundary value problem from laminar flame theory*, Arch. Rational Mech. Anal., 13 (1963), pp. 46-54.
[8] W. E. JOHNSON AND W. NACHBAR, *Laminar flame theory and the steady, linear burning of a monopropellant*, Arch. Rational Mech. Anal., 12 (1963), pp. 58-92.
[9] G. JOULIN AND P. CLAVIN, *Analyse asymptotique des conditions d'extinction des flammes laminaires*, Acta Astronautica, 3 (1976), pp. 223-240.
[10] Y. I. KANEL, *On the stabilization of solutions of the Cauchy problem for the equations arising in the theory of combustion*, Mat. Sbornik, 59 (1962), pp. 245-288.
[11] H. B. KELLER, *Numerical solution of bifurcation and nonlinear eigenvalue problems*, in Applications of Bifurcation Theory, P. Rabinowitz, ed., Academic Press, New York, 1977, pp. 359-384.
[12] B. LARROUTUROU, *Etude mathematique et modélisation numérique de phénomènes de combustion*, Thèse, Université Paris-Nord, 1987.
[13] M. MARION, *Sur les equations de flamme laminaire sans température d'ignition*, Thèse, Université Paris 6, 1983.
[14] M. H. PROTTER AND H. F. WEINBERGER, *Maximum principles in differential equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.
[15] W. C. RHEINBOLDT, *Solution fields of nonlinear equations and continuation methods*, SIAM J. Numer. Anal., 17 (1980), pp. 221-237.
[16] F. A. WILLIAMS, *Combustion Theory*, Second ed., Benjamin-Cummings, Menlo Park, CA, 1985.
[17] Y. B. ZELDOVITCH, *On the quiet flame propagation*, Z. Eksp. Teor. Fiz., 11 (1941), pp. 159-175. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1326-1334, September 1990
© 1990 Society for Industrial and Applied Mathematics
013

BOUNDED SOLUTIONS OF $\Delta u+|u|^{p-1}u-|u|^{q-1}u=0$ IN THE SUPERCRITICAL CASE*

WILLIAM C. TROY†

Abstract. The existence of bounded radial solutions of $\Delta u+|u|^{p-1}u-|u|^{q-1}u=0$ in the supercritical case $q>p>(N+2)/(N-2)$ with $N>2$ is investigated. Of particular interest are solutions that decay to zero at the rate $O(|x|^{-(N-2)})$ as $|x|\to\infty$. For each integer $J\geqq 1$ it is shown that there is at least one radially symmetric solution that decays to zero at the rate $O(|x|^{-(N-2)})$ and that has exactly J positive zeros in the interval $0<|x|<\infty$.

Key words. differential equations

AMS(MOS) subject classification. 34

1. Introduction. We investigate the existence of bounded solutions of the nonlinear elliptic problem

$$\Delta u+|u|^{p-1}u-|u|^{q-1}u=0 \quad \text{in } \mathbb{R}^N, \tag{1.1}$$

$$u\to 0 \quad \text{as } |x|\to\infty. \tag{1.2}$$

Here $x\in R^N$ denotes the independent variable, $N>2$, and $q>p>(N+2)/(N-2)$. We focus our attention on radial solutions of (1.1), (1.2). Setting $r=|x|$, we find that (1.1), (1.2) becomes

$$u''+\frac{(N-1)}{r}u'+|u|^{p-1}u-|u|^{q-1}u=0, \tag{1.3}$$

$$u(r)\to 0 \quad \text{as } r\to\infty. \tag{1.4}$$

The requirement that u be bounded in the interval $0\leqq r<\infty$ imposes the additional constraint

$$u'(0)=0. \tag{1.5}$$

The existence of bounded solutions of (1.3)-(1.5) has been thoroughly studied in the subcritical case

$$1<p<\frac{N+2}{N-2}, \qquad q=1, \quad N>1. \tag{1.6}$$

For example, Kwong [2] has proved that there is a unique positive solution. Subsequently, Jones and Küpper [1], and K. McLeod, Troy, and Weissler [4] have shown that there are infinitely many additional solutions. They show that for each integer $J\geqq 1$ there exists at least one solution of (1.3)-(1.6) that has exactly J positive zeros in the range $0<r<\infty$. The uniqueness of these solutions relative to the number of zeros in $(0,\infty)$ remains an open question. For the supercritical case $p>(N+2)/(N-2)$, $q=1$ and $N>1$ a Pohozaev argument shows that there are no solutions of (1.3)-(1.5). However, if q exceeds p, then solutions of the problem do exist. Recently, Merle and

* Received by the editors May 19, 1989; accepted for publication (in revised form) November 13, 1989. This research was partially supported by National Science Foundation grant DMS-8501531.

† Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

Peletier [5] have investigated the behavior of positive solutions of (1.3)-(1.5) for the parameter range

$$q > p > \frac{N+2}{N-2}, \qquad N > 2. \tag{1.7}$$

Their analysis of the full problem (1.1), (1.2), (1.7) shows that the decay of the radial solutions must satisfy the further constraint

$$u = O(r^{2-N}) \quad \text{as } r \to \infty. \tag{1.8}$$

They conjecture that the positive solution of the problem (1.3)-(1.5), (1.7), (1.8) is unique. Recently, Kwong et al. [3] have resolved this conjecture and have proved that the positive solution is indeed unique.

In this paper we continue the analysis described above for the supercritical case (1.7). We show that there are infinitely many additional solutions of (1.3) that satisfy (1.5), (1.7), and (1.8). To do this we follow [5] and "turn the problem around" by introducing the Emden transformation

$$t = \left(\frac{N-2}{r}\right)^{N-2}, \qquad y(t) = u(r). \tag{1.9}$$

Then (1.3)-(1.5), (1.8) is transformed into the equivalent problem

$$y'' + \frac{1}{t^k}(|y|^{p-1}y - |y|^{q-1}y) = 0, \tag{1.10}$$

$$y(0) = 0, \qquad y'(0) > 0, \tag{1.11}$$

and

$$-1 \leqq \lim_{t\to\infty} y(t) \leqq 1 \tag{1.12}$$

where $k = (2N-2)/(N-2)$.

The uniqueness theorem proved in [3] is summarized in Theorem A.

THEOREM A. *There exists a unique value $\alpha_0 > 0$ such that the solution of* (1.10) *with $y(0) = 0$ and $y'(0) = \alpha_0$ satisfies $y'(t) > 0$ for all $t > 0$ and $0 < \lim_{t\to\infty} y(t) \leqq 1$.*

We extend the results of Theorem A by seeking solutions of (1.10)-(1.12), which are not necessarily positive on the entire interval $0 < t < \infty$. A precise statement of our results is given in the following theorem.

THEOREM B. *For each integer $J \geqq 1$ there exists a value $\alpha_J > 0$ such that if the solution of* (1.10) *satisfies $(y(0), y'(0)) = (0, \alpha_J)$ then* (1.12) *holds, and the solution has exactly J isolated zeros in the interval $0 < t < \infty$.*

It remains an open problem to determine whether the solutions found in Theorem B are unique relative to the number of zeros in $0 < t < \infty$.

In the course of these investigations an extensive amount of numerical experimentation was done. All computations were done using the Adams method. It was found for $N = 3$, $p = 7$, and $q = 9$, that $\alpha_0 \approx 3.4184$, $\alpha_1 \approx 32.37$, $\alpha_2 \approx 99.8247$, and $\alpha_3 \approx 225.25$. These solutions are shown in Figs. 1-4. Our numerical work reveals the following additional information:

(i) If $0 < \alpha < \alpha_1$, then y increases until y reaches 1. Subsequently, y exceeds 1 and continues to increase throughout the rest of its interval of existence;

(ii) If $a_1 < \alpha < \alpha_2$, then y increases until $y' = 0$, then y decreases until y reaches -1, and then continues to decrease throughout the rest of its interval of existence;

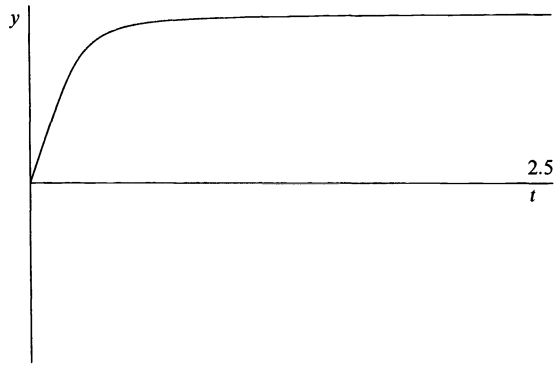


FIG. 1. *Solution with no positive zero*: $(y(0), y'(0)) \approx (0, 3.4184)$, $0 \leqq t \leqq 2.5$.

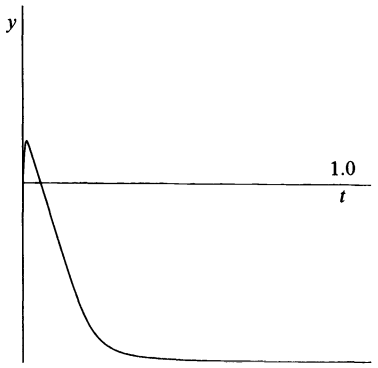


FIG. 2. *Solution with one zero*: $(y(0), y'(0)) \approx (0, 32.27)$, $0 \leqq t \leqq 1.00$.

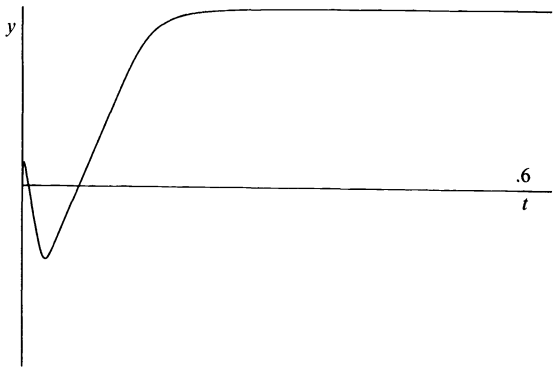


FIG. 3. *Solution with two zeros*: $(y(0), y'(0)) \approx (0, 99.8247)$, $0 \leqq t \leqq .6$.

(iii) If $\alpha_2 < \alpha < \alpha_3$, then y has two positive zeros. After the second zero, y increases until $y = 1$, then y continues to increase throughout the rest of its interval of existence. We show in our proof of Theorem B that as α continues to increase, the behavior described in (i)-(iii) persists and there are critical values of α at which there is a transition between the number of zeros of $y(t)$. At these critical values we prove that the boundary conditions at infinity hold. That is, $y' \to 0$ as $t \to \infty$, and $-1 \leqq y(\infty) \leqq 1$. Thus our method of proof consists of a topological shooting argument that uses the numerical results described above as a guide.

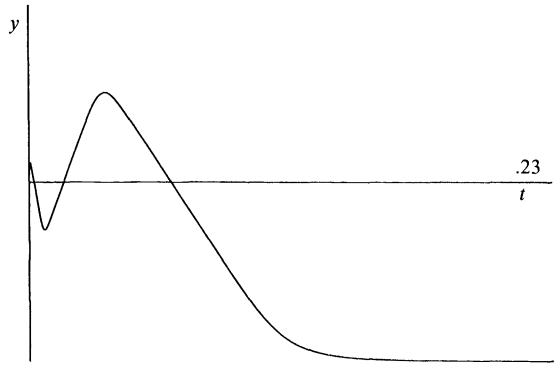


FIG. 4. *Solution with three zeros*: $(y(0), y'(0)) \approx (0, 225.25)$, $0 \leqq t \leqq .23$.

2. Proof of Theorem B. For the sake of mathematical simplicity we assume throughout that q and p are odd positive integers. (The details of the proof are exactly the same for all other choices of p and q.) Thus, with this assumption, the form of the initial value problem simplifies to

$$\frac{d^2y}{dt^2}+\frac{1}{t^k}(y^p-y^q)=0, \tag{2.1}$$

$$y(0)=0, \qquad \frac{dy}{dt}(0)=\alpha>0. \tag{2.2}$$

We let $y(t, \alpha)$ denote the solution of (2.1), (2.2). For ease of notation we will suppress the dependence of y on α whenever appropriate. Our first step is to show that if $\alpha>0$ is sufficiently large, then the solution of (2.1), (2.2) changes sign a prescribed number of times. To do this we use a scaling argument. Let

$$T=\alpha^2 t \quad \text{and} \quad y=\alpha^M Y \tag{2.3}$$

where

$$L\equiv\frac{(p-1)}{p+1-k} \quad \text{and} \quad M\equiv 1-L.$$

Then the problem (2.1), (2.2) becomes

$$\frac{d^2Y}{dT^2}+\frac{1}{T^k}(Y^p-\alpha^G Y^q)=0, \tag{2.4}$$

$$Y(0)=0, \qquad \frac{dY}{dT}(0)=1 \tag{2.5}$$

where $G=(k-2)(p-q)/(p+1-k)$. We note that $G<0$ since $p<q$, $k>2$ and $p+1-k\geqq 2/(N-2)>0$. Thus, over a given compact interval, if α is large and Y bounded, then the term $\alpha^G Y^q$ plays an insignificant role in the behavior of solutions of (2.4). This leads us to analyze the simpler problem

$$\frac{d^2Y}{dT^2}+\frac{Y^p}{T^k}=0, \tag{2.6}$$

$$Y(0)=0, \qquad \frac{dY}{dT}(0)=1. \tag{2.7}$$

We let $Y_0(T)$ denote the solution of (2.5), (2.6) and show that $Y_0(T)$ has an infinite number of zeros in the range $0<T<\infty$. To prove this we need the following technical lemma, which is similar to Lemma 1 in [3].

LEMMA 1. *Let* $T_0 \geqq 0$ *and* $\lambda > 0$. *If a solution of* (2.5), (2.6) *satisfies* $Y(T_0) = 0$ *and* $Y'(T_0) = \lambda$, *then there is a first* $T_1 > T_0$ *for which* $Y(T_1) = 0$, *and* $Y'(T_1) < 0$.

Proof. If we show that there is a first $\bar{T} > T_0$ for which $Y'(\bar{T}) = 0$, then $Y''(\bar{T}) < 0$. Subsequently, (2.5) implies that $Y'' < 0$ for $T > \bar{T}$ as long as $Y > 0$. Thus, since Y is concave down it follows that T_1 exists and that $Y'(T_1) \leqq 0$. The uniqueness of the solution $Y \equiv 0$ of (2.5) guarantees that $Y'(T_1) < 0$. Therefore it remains to show that $\bar{T}$ exists. Consider the interval $[T_0, T_0+1]$. Suppose that $Y' > 0$ for all $T \in [T_0, T_0+1]$. Then there exists a value $\eta \in (0, 1)$ such that

$$Y(T_0+1) - \eta(T_0+1)\,Y'(T_0+1) > 0. \tag{2.8}$$

From (2.8) and (2.5) it follows that $(Y - \eta TY')' = (1-\eta)\,Y' - \eta Y'' > 0$ and so

$$Y > \eta TY' \tag{2.9}$$

for $T > T_0+1$ as long as $Y' > 0$. Substituting (2.9) into (2.6), we obtain

$$Y'' \leqq -\eta^p T^{p-k} (Y')^p. \tag{2.10}$$

It then follows that $(Y')^{-p} Y'' \leqq -\eta^p T^{p-k}$ and a subsequent integration leads to

$$(Y')^{1-p} \geqq (Y'(T_0+1)^{1-p}) + \frac{\eta^p (p-1)}{(p-k+1)} (T^{p-k+1} - (T_0+1)^{p-k+1}) \tag{2.11}$$

for $T \geqq T_0+1$. Let $\tilde{T} \geqq (T_0+1)(2^{1/(p-k+1)})$. Suppose that $Y' > 0$ for all $T \in [T_0, \tilde{T}]$. Then from (2.11) we obtain

$$(Y')^{1-p} \geqq \frac{\eta^p (p-1)}{2(p-k+1)} T^{p-k+1} \tag{2.12}$$

for $T \geqq \tilde{T}$. Let $M^{p-1} = 2(p-k+1)/(\eta^p (p-1))$. Then (2.12) simplifies to

$$Y' \leqq M T^{(k-p-1)/(p-1)}. \tag{2.13}$$

Note that $-1 < (k-p-1)/(p-1) < 0$ since $N > 2$, $k = (2N-2)/(N-2)$ and $p > (N+2)/(N-2)$. An integration of (2.13) leads to

$$Y(T) \leqq Y(\tilde{T}) + LT^{((k-2)/(p-1))} \tag{2.14}$$

where $L = (p-1)M/(k-2)$. We use the estimates (2.13) and (2.14) to obtain our result. Suppose, for the sake of contradiction, that $Y' > 0$ for all $T > T_0$. Then (2.13) and (2.14) hold for all $T \geqq \tilde{T}$. We multiply (2.6) by Y and $TY'(T)$, integrate, and obtain the equations

$$Y(T)\,Y'(T) - \int_{T_0}^{T} (Y')^2\,dx = -\int_{T_0}^{T} \frac{Y^{p+1}(x)}{x^k}\,dx \tag{2.15}$$

and

$$\begin{aligned} &\frac{-T^{1-k}}{p+1} (Y(T))^{p+1} + \int_{T_0}^{T} \frac{(1-k)}{(p+1)} \frac{Y^{p+1}(x)}{x^k}\,dx \\ &\quad = -\frac{1}{2}\int_{T_0}^{T} (Y'(x))^2\,dx + \frac{T(Y'(T))^2}{2} - \frac{T_0(Y'(T_0))^2}{2}. \end{aligned} \tag{2.16}$$

We combine (2.15) and (2.16) into

$$\frac{T}{2}(Y'(T))^2+\frac{T^{1-k}Y^{p+1}}{p+1}(T)-\frac{Y(T)}{2}Y'(T)-\frac{T_0(Y'(T_0))^2}{2} \tag{2.17}$$
$$=\int_{T_0}^{T}\left(\frac{1}{2}+\frac{1-k}{p+1}\right)x^{-k}Y^{p+1}(x)\,dx.$$

It follows from the definitions of p and k, and (2.14) that

$$\int_{T_0}^{\infty}\left(\frac{1}{2}+\frac{1-k}{p+1}\right)x^{-k}Y^{p+1}(x)\,dx$$

converges to a positive value $\rho>0$. This and (2.17) imply that

$$\frac{T(Y'(T))^2}{2}+\frac{T^{1-k}}{p+1}(Y(T))^{(p+1)}-\frac{Y(T)Y'(T)}{2}-\frac{T_0(Y'(T_0))^2}{2}\geqq\frac{\rho}{2}>0 \tag{2.18}$$

for all large T. However (2.13) and (2.14) imply that $T(Y'(T))^2/2\to 0$, $Y(T)Y'(T)\to 0$ and $T^{1-k}(Y(T))^{p+1}\to 0$ as $T\to\infty$. This leads us to conclude that

$$\overline{\lim_{T\to\infty}}\left(\frac{T(Y'(T))^2}{2}+\frac{T^{1-k}}{p+1}(Y(T))^{(p+1)}-\frac{Y(T)Y'(T)}{2}-\frac{T_0(Y'(T_0))^2}{2}\right)\leqq 0$$

contradicting (2.18).

This completes the proof.

We use Lemma 1 to show that $Y_0(T)$ oscillates infinitely often as $T\to\infty$. A precise statement is given in the next lemma.

LEMMA 2. *There is an unbounded, increasing sequence* $\{T_i\}_{i\geqq 1}$ *of positive values such that*

(i) $Y_0(T_i)=0$ *and* $Y_0'(T_i)\neq 0$ *for each* $i\geqq 1$,

(ii) $Y_0(T)>0$ *for all* $T\in(0,T_1)$, $Y_0(T_1)=0$, *and*

(iii) $Y_0(T)\neq 0$ *if* $T\in\{T_1,T_2,\cdots\}$.

Proof. It follows from Lemma 1 that T_1 exists satisfying (ii), and that $Y_0'(T_1)<0$. Since p is odd, $g(T)\equiv -Y_0(T)$ satisfies (2.5), $g(T_1)=0$, and $g'(T_1)>0$. This and Lemma 1 imply that there is a first $T_2>T_1$ such that $g(T_2)=0$, and $g'(T_2)<0$. Thus $Y_0(T)<0$ for all $T\in(T_1,T_2)$, $Y_0(T_2)=0$ and $Y_0'(T_2)>0$. Repeated use of Lemma 1 and a mathematical induction argument lead us to conclude that there are values $T_2<T_3<T_4<\cdots$ that satisfy (i) and (iii), and the lemma follows.

Next, we proceed with the proof of Theorem B and construct our "shooting" sets. For each integer $J\geqq 1$ we define $A_J=\{\hat{\alpha}>0 \mid \text{if } \alpha>\hat{\alpha}, \text{ then the solution of (2.1), (2.2) has at least } J+1 \text{ zeros in } (0,\infty)\}$.

LEMMA 3. *For each* $J\geqq 1$ *the set* A_J *is open, nonempty and* $\inf A>0$.

Proof. It follows from Theorem A that there is an α_0 such that $y'(t,\alpha_0)>0$ for all $t>0$ and $0<y(\infty,\alpha_0)<1$. Thus α_0 is a lower positive bound for each A_J. Next, let $\hat{\alpha}\in A_J$ and let $\hat{t}$ denote a positive zero of the solution of (2.1) satisfying $y(0)=0$ and $y'(0)=\hat{\alpha}$. Uniqueness of solutions guarantees that $y'(\hat{t})\neq 0$. This and continuity imply that $\alpha\in A_J$ if $|\alpha-\hat{\alpha}|$ is sufficiently small. Thus each A_J is open. It remains to show that each A_J is nonempty. Recall that the solution $Y_0(T)$ of (2.5), (2.6) oscillates infinitely often as $T\to\infty$. Thus there is an interval $I=[0,\bar{\nu}]$, $\bar{\nu}<\infty$, such that $Y_0(T)$ has at least $J+2$ positive zeros in I. Furthermore, $Y_0(T)$ is bounded on I since I is compact. Thus the solution $Y(T)$ of (2.4), (2.5) converges uniformly to $Y_0(T)$ on the interval I as $\alpha\to\infty$. From this and the transformation (2.2) it follows that the solution

of (2.5), (2.6) has at least $J+1$ positive zeros in I for all large α. Thus each A_J is nonempty and the proof is complete.

We need two further technical lemmas to complete our proof.

LEMMA 4. *There exists $\delta>0$ such that if a solution of* (2.1) *satisfies $y(\hat{t})=0$ and $|y'(\hat{t})|\leqq\delta$ for some $\hat{t}\geqq 0$ then there is a $\tilde{t}>\hat{t}$ such that*

(i) $|y'(t)|\neq 0$ *for all* $t\in[\hat{t},\tilde{t}]$,

(ii) $|y(\tilde{t})|=1$, *and*

(iii) $|y'(t)|\neq 0$ *for* $t\geqq\tilde{t}$ *as long as the solution exists.*

Proof. Let $y'(\hat{t})=\lambda$ and assume that $\lambda>0$. The details of the proof are exactly the same if $\lambda<0$ and are therefore omitted for the sake of brevity. For $t\geqq\hat{t}$, as long as $y<1$ it follows from (2.1) that $y''<0$ so that $y<\lambda(t-\hat{t})$. From this and (2.1) we conclude that $y''\geqq -y^p/t^k\geqq-\lambda^p(t-\hat{t})^{p-k}$. Integrating, we obtain

$$y'\geqq\lambda-\frac{\lambda^p(t-\hat{t})^{p-k+1}}{(p-k+1)} \tag{2.19}$$

and

$$y\geqq\lambda(t-\hat{t})-\frac{\lambda^p(t-\hat{t})^{p-k+2}}{(p-k+1)(p-k+2)}. \tag{2.20}$$

We assume that

$$0<\lambda<\min\left\{(p-k+1)^{1/(k-2)},\left(\frac{1}{2(p-k+1)(p-k+2)}\right)^{1/(k-2)}\right\}.$$

It then follows from (2.19) that

$$\lambda-\frac{\lambda^{k-1}}{p-k+1}\leqq y'<\lambda \tag{2.21}$$

for all $t\in[\hat{t},\hat{t}+1/\lambda]$. Thus we are assured that $y'>0$ for all $t\in[\hat{t},\hat{t}+1/\lambda]$ since $0<\lambda<(p-k+1)^{1/(k-2)}$. Furthermore, from (2.20) and the restrictions on λ it follows that

$$\frac{1}{2}\leqq 1-\frac{\lambda^{k-2}}{(p-k+1)(p-k+2)}\leqq y\left(\hat{t}+\frac{1}{\lambda}\right)<1. \tag{2.22}$$

Next, we define the function $g(y)=1-y^{q-p}$ and set $D=\sup_{(1/2)\leqq y\leqq 1}|g'(y)|$. It follows from the mean value theorem that

$$g(y)\leqq D(1-y)\quad\forall y\in[\tfrac{1}{2},1]. \tag{2.23}$$

Thus (2.1), (2.22), and (2.23) imply that

$$y''\geqq-\gamma\lambda^{k-2}t^{-k} \tag{2.24}$$

where $\gamma=D/((p-k+1)(p-k+2))$ for $t\geqq\hat{t}+1/\lambda$ as long as $y'>0$ and $y\leqq 1$. From (2.21) and (2.24) we obtain

$$\begin{aligned} y'&\geqq\lambda-\frac{\lambda^{k-1}}{(p-k+1)}-\frac{\gamma\lambda^{k-2}}{(1-k)}(t^{1-k}-(\hat{t}+\lambda^{-1})^{1-k})\\ &\geqq\lambda-\frac{\lambda^{k-1}}{(p-k+1)}-\frac{\gamma\lambda^{2k-3}}{(k-1)}. \end{aligned} \tag{2.25}$$

Now let $\delta > 0$ satisfy

$$1 - \frac{\lambda^{k-2}}{(p-k+1)} - \frac{\gamma\lambda^{2k-4}}{(k-1)} > \frac{1}{2} \tag{2.26}$$

for all $\lambda \in (0, \delta)$. From (2.25) and (2.26) we conclude that $y' \geqq \lambda/2$ for $t \geqq \hat{t} + 1/\lambda$ as long as $y \leqq 1$. An integration shows that there is a first $\tilde{t} > \hat{t}$ for which $y(\tilde{t}) = 1$ and $y'(\tilde{t}) \geqq 0$. Uniqueness of solutions implies that $y'(\tilde{t}) > 0$. Finally, it follows from (2.1) that $y'' > 0$, $y' > 0$, and $y > 1$ for $t \geqq \tilde{t}$ as long as the solution is defined. This completes the proof of the lemma.

LEMMA 5. *Let $\bar{\alpha} > 0$ such that for some $\bar{T} > 0$, $|y'(t, \bar{\alpha})| \neq 0$ for all $t \geqq \bar{T}$, and $\lim_{t\to\infty} (y(t, \bar{\alpha}), y'(t, \bar{\alpha})) = (\bar{y}, 0)$, where $0 < |\bar{y}| \leqq 1$. Suppose that $y(t, \bar{\alpha})$ has exactly $m \geqq 0$ zeros in $(0, \infty)$. If $|\alpha - \bar{\alpha}| > 0$ is sufficiently small, then $y(t, \alpha)$ has at most $m+1$ positive zeros in its maximal interval of existence.*

Proof. Suppose, first of all, that $y(t, \bar{\alpha}) > 0$ and $y'(t, \bar{\alpha}) > 0$ for all $t > \bar{T}$, and that $0 < \bar{y} \leqq 1$. Define the values

$$l = \sup_{0 \leqq y \leqq 1} |y^p - y^q| \quad \text{and} \quad \hat{t} = \max\{\bar{T}, (l/(\delta(k-1)))^{1/(k-1)}\} \tag{2.27}$$

where δ satisfies Lemma 4. It follows from continuity that $y(t, \alpha)$ has at most m zeros in $(0, \hat{t}]$, $y(\hat{t}, \alpha) > 0$, and $y'(\hat{t}, \alpha) > 0$, if $|\alpha - \bar{\alpha}| > 0$ is sufficiently small. Suppose that $y(t^*, \alpha) = 0$ for some first $t^* > \hat{t}$. Then it follows that there exists $\bar{t} \in (\hat{t}, t^*)$ such that $y'(\bar{t}, \alpha) = 0$, and $y''(t, \alpha) < 0$ for $\bar{t} < t < t^*$. Furthermore, from (2.1) and (2.27) we conclude that $y'' > -lt^{-k}$ for $t \in (\bar{t}, t^*)$. An integration leads to $y' > -l/(1-k)(t^{1-k} - \bar{t}^{1-k})$ for $\bar{t} \leqq t \leqq t^*$. From this and (2.27) we conclude that $y(t^*, \alpha) = 0$ and $-\delta < y'(t^*, \alpha) < 0$. But then Lemma 4 implies that $y'(t, \alpha) < 0$ for $t > t^*$ as long as the solution exists. The details for the case $-1 \leqq \bar{y} < 0$ are similar and are omitted for the sake of brevity. This completes the proof of the lemma.

We now proceed with the final details of proving our theorem. For each $J \geqq 1$ we define $\alpha_J = \inf A_J$. We claim that the solution $y(t, \alpha_J)$ satisfies the theorem. That is, $y(t, \alpha_J)$ has exactly J zeros in $(0, \infty)$, and $0 < |y(\infty, \alpha_J)| \leqq 1$. We let $(0, \nu_J)$ denote the maximal interval of existence of $y(t, \alpha_J)$. There are three possibilities to consider. First, suppose that $y(t, \alpha_J)$ has at most $J-1$ zeros in $(0, \nu_J)$. Then there exists a value $a \in (0, \nu_J)$ such that $y(t, \alpha_J) \neq 0$ for all $t \in [a, \nu_J)$. Without loss of generality we may assume that $y(t, \alpha_J) > 0$ for all $t \in [a, \nu_J)$ since $g \equiv -y(t, \alpha_J)$ is also a solution of (2.1). If $y'(\bar{t}, \alpha_J) \leqq 0$ for some $\bar{t} \in (a, \nu_J)$ then (2.1) implies that $y'(t, \alpha_J) < 0$ and $y''(t, \alpha_J) < 0$ for $t > \bar{t}$ until $y(t, \alpha_J)$ equals zero, a contradiction. Therefore $y'(t, \alpha_J) > 0$ for all $t \in [a, \nu_J)$. Suppose that $y(\tilde{t}, \alpha_J) = 1$ at some $\tilde{t} \in [a, \nu_J)$. Then (2.1) implies that $y''(t, \alpha_J) > 0$, $y'(t, \alpha_J) > 0$, and $y(t, \alpha_J) > 1$ for all $t \in (\tilde{t}, \nu_J)$. It follows from continuity that if $|\alpha - \alpha_J|$ is small, then there is a value $\tilde{t} = \tilde{t}(\alpha) > 0$ such that $y(t, \alpha)$ has at most $J-1$ zeros in $(0, \tilde{t})$, $y(\tilde{t}, a) = 1$ and $y'(\tilde{t}, \alpha) > 0$. Again (2.1) implies that $y''(t, \alpha) > 0$, $y'(t, \alpha) > 0$, and $y(t, a) > 1$ for $t > \tilde{t}$ as long as the solution exists. Therefore $\alpha \notin A_J$ if $|\alpha - \alpha_J|$ is sufficiently small, contradicting the definition of α_J. We conclude that $y'(t, \alpha_J) > 0$ and $0 < y(t, \alpha_J) < 1$ for all $t \in (a, \nu_J)$. If $\nu_J < \infty$, then one of y or y' becomes unbounded as $t \to \nu_J$. However, it follows from (2.1) that $y''(t, \alpha) < 0$ for all $t \in (a, \nu_J)$ so that $y'(t, \alpha_J)$ and $y(t, \alpha_J)$ are bounded. Therefore $\nu_J = \infty$ and it follows that $\lim_{t\to\infty} (y(t, \alpha_J), y'(t, \alpha_J)) = (\bar{y}, 0)$ for some $\bar{y} \in (0, 1]$. But this and Lemma 5 show that $y(t, \alpha)$ has at most J zeros if $\alpha - \alpha_J > 0$ is sufficiently small. This contradicts the definition of α_J since A_J is open. This leads us to consider our second possibility, that $y(t, \alpha_J)$ has at least $J+1$ zeros in $(0, \nu_J)$. However, it follows from continuity that if $\alpha - a_J > 0$ is sufficiently small, then $y(t, \alpha)$ has at least $J+1$ zeros in $(0, \nu_J)$, again

contradicting the definition of α_J. Therefore, it must be the case that $y(t, \alpha_J)$ has exactly J zeros in $(0, \nu_J)$. Thus there exists a value $a \in (0, \nu_J)$ such that $y(t, \alpha_J) \neq 0$ on (a, ν_J) and we again may assume that $y(t, \alpha_J) > 0$ for all $t \in (\alpha, \nu_J)$. Suppose that $y(\tilde{t}, \alpha_J) = 1$ at some first $\tilde{t} \in (0, \nu_J)$. Equation (2.1) implies that $y''(t, \alpha_J) > 0$ and $y'(t, \alpha_J) > 0$ for all $t \in (\tilde{t}, \nu_J)$. It follows from continuity that if $\alpha - a_J > 0$ is sufficiently small then there exists a first $\tilde{t} = \tilde{t}(\alpha) > 0$ such that $y(\tilde{t}, \alpha) = 1$ and $y(t, \alpha)$ has exactly J zeros in $(0, \tilde{t})$. Subsequently, (2.1) implies that $y''(t, \alpha) > 0$ and $y'(t, \alpha) > 0$ for $t > \tilde{t}$ as long as the solution $y(t, \alpha)$ exists. Thus $\alpha \notin A_J$ if $\alpha - \alpha_J > 0$ is sufficiently small, contradicting the definition of α_J since A_J is open. Therefore it must be the case that $y'(t, \alpha_J) > 0$ and $0 < y(t, \alpha_J) < 1$ for all $t \in (0, \nu_J)$. As shown in case (i) above since $y(t, \alpha_J)$ is uniformly bounded on $[0, \nu_J)$ then also $y''(t, \alpha)$ is uniformly bounded on $[0, \nu_J)$ and it must be the case that $\nu_J = \infty$. Thus we conclude that $y'(t, \alpha_J) > 0$ for all $t \geqq a$ and that $0 < \lim_{t \to \infty} y(t, \alpha_J) \leqq 1$. This completes the proof.

REFERENCES

[1] C. JONES AND T. KÜPPER, *On the infinitely many solutions of a semilinear elliptic equation*, SIAM J. Math. Anal., 17 (1986), pp. 803–835.

[2] M. K. KWONG, *Uniqueness of positive solutions of* $\Delta u + u^p - u = 0$ *in* $\mathbb{R}^N$, Arch. Rational Mech., 105 (1989), pp. 243–266.

[3] M. K. KWONG, J. B. MCLEOD, L. A. PELETIER, AND W. C. TROY, *Uniqueness of a ground state of* $\Delta u + u^p - u^q = 0$, J. Differential Equations, submitted.

[4] M. MCLEOD, W. C. TROY, AND F. B. WEISSLER, *Radial solutions of* $\Delta u + f(u) = 0$ *with prescribed numbers of zeros*, J. Differential Equations, 83 (1990), pp. 368–378.

[5] F. MERLE AND L. A. PELETIER, *Asymptotic behavior of positive solutions of elliptic equations with critical and supercritical growth* I. *The radial case*, preprint.

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1335-1348, September 1990
© 1990 Society for Industrial and Applied Mathematics
014

ON A RECURRENCE FORMULA ASSOCIATED WITH STRONG DISTRIBUTIONS*

A. SRI RANGA†

Abstract. Polynomials satisfying a certain three-term recurrence relation are studied. The properties of these polynomials and an associated strong distribution are observed under various conditions on the coefficients of the recurrence relation. Some examples are given to illustrate these results.

Key words. three-term recurrence relation (or formula), $\hat{J}$-fractions, strong distribution functions, Stieltjes functions

AMS(MOS) subject classifications. 11A55, 42C05, 30C15, 40A15

1. Introduction.

The recurrence formula or relation in our study takes the form

$$\begin{aligned} Q_{2n-1}(z) &= \{(1+\alpha_{2n-1})z-\beta_{2n-1}\}Q_{2n-2}(z)-\alpha_{2n-1}z^2Q_{2n-3}(z), \\ Q_{2n}(z) &= (z-\beta_{2n})Q_{2n-1}(z)-\alpha_{2n}Q_{2n-2}(z), \end{aligned} \qquad n\geqq 1, \tag{1.1}$$

where

$$\begin{aligned} &Q_{-1}(z)=0, \qquad Q_0(z)=1, \\ &\alpha_1=0, \quad \alpha_{n+1}>0, \quad \beta_n\in\mathbb{R}, \quad n\geqq 1. \end{aligned} \tag{1.2}$$

It can easily be verified that, for any $n\geqq 0$, $Q_n(z)$ is a monic polynomial of degree n, satisfying in particular

$$Q_n(z)=z^n+\cdots+q_1^{(n)}z+q_0^{(n)},$$

where

$$\begin{aligned} &q_0^{(0)}=1, \quad q_0^{(1)}=-\beta_1, \quad q_0^{(2n)}=\prod_{r=1}^{n}(\beta_{2r-1}\beta_{2r}-\alpha_{2r}), \\ &q_0^{(2n+1)}=-\beta_{2n+1}q_0^{(2n)}, \\ &q_1^{(2n+1)}=(1+\alpha_{2n+1})q_0^{(2n)}-\beta_{2n+1}q_1^{(2n)} \end{aligned} \tag{1.3}$$

for all $n\geqq 1$. From the theory of continued fractions we also have that $Q_n(z)$ is the denominator of the nth convergent of the fraction

$$\frac{\alpha_1}{z-\beta_1}-\frac{\alpha_2}{z-\beta_2}-\frac{\alpha_3z^2}{(1+\alpha_3)z-\beta_3}-\frac{\alpha_4}{z-\beta_4}-\frac{\alpha_5z^2}{(1+\alpha_5)z-\beta_5}-\cdots, \tag{1.4}$$

which is called a regular $\hat{J}$-fraction.

So far, in all the attempts at resolving the so-called strong Hamburger moment problem using regular continued fractions, the two regular $\hat{J}$-fractions, one being (1.4) and the other being of the form

$$\frac{\alpha_1}{z-\beta_1}-\frac{\alpha_2z^2}{(1+\alpha_2)z-\beta_2}-\frac{\alpha_3}{z-\beta_3}-\frac{\alpha_4z^2}{(1+\alpha_4)z-\beta_4}-\frac{\alpha_5}{z-\beta_5}-\cdots,$$

* Received by the editors August 24, 1987; accepted for publication (in revised form) November 7, 1989. This research was supported by CNPq of Brazil.

† Instituto de Ciências Matemátics de São Carlos, Universidade de São Paulo, São Carlos, São Paulo, Brazil.

are known to provide the best partial solutions to this problem (Sri Ranga [10]). Here the term "regular" implies that there is a repetitive pattern in the functional forms taken by the partial coefficients.

In [11] Sri Ranga examines the convergence properties of a class of $\hat{J}$-fractions and gives some criteria for the convergence of a regular real $\hat{J}$-fraction. This regular $\hat{J}$-fraction, though it appears to be different from those given above, is equivalent to the fraction (1.4). Even though these two equivalent continued fractions have the same convergence behaviour, it will be seen (Theorem 3.2 in § 3) that some of the convergence criteria given in [11] are much simpler if they are written in terms of the coefficients α_n and β_n in (1.4). In view of this and because the denominators of (1.4) are monic polynomials, it may be considered that (1.4) is of a more natural form than its equivalent form given in [11].

A well-known result on three-term relations is that of the orthogonal polynomials. Associated with any positive distribution $d\psi(t)$ on $(-\infty, \infty)$ for which the moments $c_n = \int_{-\infty}^{\infty} t^n \, d\psi(t)$ exist for all $n \geqq 0$, there exists a sequence of orthogonal (monic) polynomials $\{B_n(z)\}$ satisfying a recurrence formula of the form

$$B_n(z) = (z - m_n) B_{n-1}(z) - l_n B_{n-2}(z), \qquad n \geqq 1, \tag{1.5}$$

where $B_{-1}(z) = 0$, $B_0(z) = 1$, $m_n \in \mathbb{R}$, and $l_n > 0$. A not-so-familiar result regarding this three-term relation is attributed to Favard [5]. This result can be stated as: "Given (1.5), associated with that relation there always exists a positive distribution $d\psi(t)$ such that the polynomials generated from the relation form an orthogonal sequence with respect to this distribution." However, (1.5) generally does not uniquely characterize its distribution $d\psi(t)$. For the uniqueness of $d\psi(t)$, the coefficients l_n and m_n must also satisfy further conditions. The simplest of such additional conditions for the uniqueness of $d\psi(t)$ is the boundedness of these coefficients.

In recent years considerable work has been done on extracting the properties of the associated distribution $d\psi(t)$ from three-term formulae of the form (1.5). Many interesting results have been obtained by, for example, Blumenthal [1], Chihara [2], [4], Nevai [8], [9], and Van Assche [13].

In this article we consider an analogous study on the three-term recurrence formula (1.1). We believe that the results presented here can be useful in justifying the applications of two-point Padé approximation and associated continued fractions to problems in science and engineering. One result, that of Theorem 3.2, can be of interest to those working on the moment problems, where it provides conditions for the uniqueness of a certain strong Hamburger moment problem. One interesting aspect of the classical orthogonal polynomials that may have led to the extended theories on the subject is the mathematical elegance in their recurrence relations. That is, the coefficients in their recurrence relations can be explicitly given by elegant mathematical formulae. Since the recurrence relations of (4.1)–(4.3) in § 4 exhibit such mathematical elegance, we hope that the associated polynomials will also inspire such extended theories and that they will merit a study analogous to that of the classical orthogonal polynomials.

2. The strong distribution. From the studies of the $\hat{J}$-fractions and the strong moment problems [10], it has become evident that, given a bounded positive distribution $d\psi(t)$ on $(-\infty, \infty)$, with existence of the moments $c_m = \int_{-\infty}^{\infty} t^m \, d\psi(t)$ for all values of m including negative, there exists a unique $\hat{J}$-fraction of the form (1.4) corresponding to the Stiltjes function $\int_{-\infty}^{\infty} d\psi(t)/(z-t)$, provided that the moments also satisfy

$$H_{2n}^{(-2n)} > 0, \quad H_{2n+1}^{(-2n)} > 0, \quad \text{and} \quad H_{2n}^{(-2n+1)} \neq 0, \quad n \geqq 1.$$

Here $H_n^{(k)}$ are the Hankel determinants associated with c_m. (Distributions for which the moments exist for all positive and negative values of m are also known as strong distributions.) Correspondence of this $\hat{J}$-fraction is such that the convergents are two-point (zero and infinity) Padé approximations of the Stieltjes function.

The determinantal conditions $H_{2n}^{(-2n+1)} \neq 0$, $n \geqq 1$, ensure that the denominator $Q_n(z)$ of the nth convergent of this fraction satisfies $Q_n(0) \neq 0$ when n is even. Thus looking at (1.3) we realise that, in the three-term formula (1.1) satisfied by these polynomials, the coefficients must also satisfy

$$\beta_{2n-1}\beta_{2n} - \alpha_{2n} \neq 0, \qquad n \geqq 1. \tag{2.1}$$

We now start from the three-term relation (1.1) and proceed to establish the existence of a distribution associated with it. We use in conjunction another sequence of polynomials $\{P_n(z)\}$ generated by

$$\begin{aligned} P_{2n}(z) &= (z-\beta_{2n})P_{2n-1}(z) - \alpha_{2n}P_{2n-2}(z), \\ P_{2n+1}(z) &= \{(1+\alpha_{2n+1})z - \beta_{2n+1}\}P_{2n}(z) - \alpha_{2n+1}z^2P_{2n-1}(z), \end{aligned} \qquad n \geqq 1, \tag{2.2}$$

where $P_0(z) = 0$, $P_1(z) = a_1$, and α_n, β_n are as in (1.1).

This relation differs from (1.1) only in the initial conditions of the polynomials. For any value of $a_1 \neq 0$, (2.2) generates a nontrivial sequence of polynomials $\{P_n(z)\}$. In this case we can write

$$P_n(z) = a_1 z^{n-1} + \text{lower-order terms.}$$

These polynomials are also easily verified as being the numerators of the convergence of the $\hat{J}$-fractions (1.4).

From (1.1) and (2.2) we have that the functions $T_n(z)$ and $U_n(z)$ defined by

$$T_n(z) = \{Q_n'(z)Q_{n-1}(z) - Q_{n-1}'(z)Q_n(z)\}, \qquad n \geqq 1,$$

$$U_n(z) = \{P_n(z)Q_{n-1}(z) - P_{n-1}(z)Q_n(z)\}, \qquad n \geqq 1,$$

satisfy the relations

$$T_{2n}(z) = \{Q_{2n-1}(z)\}^2 + \alpha_{2n}T_{2n-1}(z),$$

$$T_{2n+1}(z) = \{Q_{2n}(z)\}^2 + \alpha_{2n+1}\{Q_{2n}(z) - zQ_{2n-1}(z)\}^2 + \alpha_{2n+1}\alpha_{2n}z^2T_{2n-1}(z),$$

$$U_{2n}(z) = \alpha_{2n}U_{2n-1}(z),$$

$$U_{2n+1}(z) = \alpha_{2n+1}z^2U_{2n}(z)$$

for $n \geqq 1$, with $T_1(z) = 1$ and $U_1(z) = a_1$. Under (1.2) the functions $T_n(z)$, $n \geqq 1$, are hence strictly positive for all real values of z other than zero. If we also assume that the coefficients of (1.1) satisfy (2.1) (i.e., $Q_{2n}(0) \neq 0$, $n \geqq 1$), then $T_n(z)$ are also positive for $z = 0$. This enables us to establish that the roots of $Q_n(z)$ are all real, distinct, and different from those of $Q_{n-1}(z)$.

We take $a_1 > 0$. Then the functions $U_n(z)$, $n \geqq 1$, are also positive for all real values of z other than zero. For $n \geqq 3$, $U_n(z)$ takes the value zero when $z = 0$. Hence all the roots of $Q_{2n}(z)$ are different from those of $P_{2n}(z)$, and all the nonzero roots of $Q_{2n+1}(z)$ are different from those of $P_{2n+1}(z)$. It is possible that at most one of the roots of $Q_{2n+1}(z)$ is zero. For any $n \geqq 1$, when zero is a root of $Q_{2n+1}(z)$ it is also a root of $P_{2n+1}(z)$.

From these results it immediately follows that the quotient $P_n(z)/Q_n(z)$ has a partial decomposition of the form

$$\frac{P_n(z)}{Q_n(z)} = \sum_{r=1}^{n} \frac{l_r^{(n)}}{z - z_r^{(n)}}, \qquad n \geqq 1, \tag{2.3}$$

where $z_r^{(n)}$ are the roots of $Q_n(z)$ and

$$l_r^{(n)} = P_n(z_r^{(n)})/Q_n'(z_r^{(n)}), \qquad r = 1, 2, \cdots, n.$$

By noting that $l_r^{(n)}$ is also equal to $U_n(z_r^{(n)})/T_n(z_r^{(n)})$, we find that it is a positive number except in the case when n is an odd number greater than 2 and that $z_r^{(n)}$ is equal to zero. In this case $l_r^{(n)}$ is also equal to zero.

Taking the limit of $\{zP_n(z)/Q_n(z)\}$ as $z \to \infty$ in (2.3), we also obtain

$$\sum_{r=1}^{n} l_r^{(n)} = a_1.$$

Therefore, if we define a step function $\psi_n(t)$ by

$$\psi_n(t) = \begin{cases} 0, & -\infty < t \leqq z_1^{(n)} \\ \sum_{s=1}^{r} l_s^{(n)}, & z_r^{(n)} < t \leqq z_{r+1}^{(n)}, \qquad r = 1, 2, \cdots, n-1, \\ a_1, & z_n^{(n)} < t < \infty, \end{cases}$$

then from the definition of the Stieltjes integral we have

$$\frac{P_n(z)}{Q_n(z)} = \int_{-\infty}^{\infty} \frac{1}{z - t}\, d\psi_n(t), \qquad n \geqq 1. \tag{2.4}$$

We now state a result (see [10, p. 271]) on the correspondence behaviour of the quotients $P_n(z)/Q_n(z)$, $n \geqq 1$.

THEOREM 2.1. *With conditions* (1.2) *and* (2.1), $L_n(z) = P_n(z)/Q_n(z)$, $n \geqq 1$, *which are the convergents of the* $\hat{J}$*-fraction* (1.4), *correspond to two formal power series expansions*

$$f(z) = \frac{c_0}{z} + \frac{c_1}{z^2} + \frac{c_2}{z^3} + \frac{c_3}{z^4} + \cdots,$$

$$g(z) = -c_1 - c_{-2}z - c_{-3}z^2 - c_{-4}z^3 - \cdots$$

such that

$$\begin{aligned} f(z) - L_{2n}(z) &= \gamma_{2n} z^{-2n-1} + O(1/z)^{2n+2}, \\ f(z) - L_{2n+1}(z) &= \gamma_{2n+1} z^{-2n-3} + O(1/z)^{2n+4}, \end{aligned} \qquad n \geqq 0,$$

where

$$\begin{aligned} &\gamma_0 = a_1 = c_0, \qquad \gamma_{2n} = \alpha_{2n+1}\alpha_{2n} \cdots \alpha_2 a_1, \\ &\gamma_{2n+1} = (1 + \alpha_{2n+3})\alpha_{2n+2}\alpha_{2n+1} \cdots \alpha_2 a_1, \\ &Q_{2n}(z)g(z) - P_{2n}(z) = \delta_{2n} z^{2n} + O(z)^{2n+1}, \\ &Q_{2n+1}(z)g(z) - P_{2n+1}(z) = \delta_{2n+1} z^{2n} + O(z)^{2n+1}, \end{aligned} \qquad n \geqq 0,$$

where

$$\delta_{2n} = -\beta_{2n+2}\gamma_{2n}/Q_{2n+2}(0), \qquad \delta_{2n+1} = \alpha_{2n+2}\gamma_{2n}/Q_{2n+2}(0).$$

With the requirement $Q_{2n}(0) \neq 0$, the correspondence behaviour of $L_{2n}(z)$, $n \geqq 1$, to the formal series expansions $g(z)$ is very clear. Since we have not considered the behaviour of $Q_{2n+1}(0)$, the correspondence of $L_{2n+1}(z)$ to the series expansion $g(z)$ is not entirely evident. However, from (1.3) we have that, if $Q_{2n+1}(0) = -\beta_{2n+1} Q_{2n}(0) = 0$, then $Q'_{2n+1}(0) = (1+\alpha_{2n+1}) Q_{2n}(0) \neq 0$. This result ensures that $L_{2n+1}(z)$ corresponds to at least $(2n-1)$ terms of the series expansion $g(z)$ for $n \geqq 1$.

Now, by using the results of the preceding theorem with (2.4), we are able to give the next theorem.

THEOREM 2.2. *When conditions* (1.2) *and* (2.1) *hold, there always exists a function* $\psi(t)$, *bounded, nondecreasing in* $(-\infty, \infty)$, *and with infinitely many points of increase, such that*

$$\lim_{r\to\infty} \psi_{n(r)}(t) = \psi(t),$$

$$\lim_{r\to\infty} L_{n(r)}(z) = \int_{-\infty}^{\infty} d\psi(t)/(z-t),$$

$$c_m = \int_{-\infty}^{\infty} t^m \, d\psi(t), \qquad m = \cdots, -2, -1, 0, 1, 2, \cdots.$$

Here $\{n(r)\}$ *is an increasing subsequence of the sequence of positive integers.*

This theorem follows from arguments analogous to those used for the J-fraction and positive T-fraction by, respectively, Wall [14] and Jones, Thron, and Waadeland [7].

3. The $\psi(t)$ and the polynomials $Q_n(z)$ and $P_n(z)$. Theorem 2.2 states that there exists a distribution function $\psi(t)$ such that its Stieltjes function

$$S(\psi(t); z) = \int_{-\infty}^{\infty} d\psi(t)/(z-t)$$

is a limit of a subsequence of $\{L_n(z)\}$. As a consequence, $S(\psi(t); z)$ has formal expansions $f(z)$ and $g(z)$. There may also be other distribution functions whose Stieltjes functions are limits of other subsequences of $\{L_n(z)\}$ (or even limits of other similar sequences) and hence have the same formal expansions. That is, the associated distribution function may not be unique. (We recall that all distribution functions differing from each other only by constant values are nondistinct.) At the end of this section we provide conditions on the coefficients α_n and β_n that ensure the uniqueness of $\psi(t)$.

For now, let $\psi(t)$ be any distribution function such that $S(\psi(t); z)$ has formal expansions $f(z)$ and $g(z)$. Then we have Theorem 3.1 below.

THEOREM 3.1. *The polynomials* $Q_n(z)$ *and* $P_n(z)$ *can be given in terms of* $\psi(t)$ *as*

$$\int_{-\infty}^{\infty} t^{-2n+s} Q_{2n}(t) \, d\psi(t) = \begin{cases} 0, & 0 \leqq s \leqq 2n-1, \\ \gamma_{2n}, & s = 2n, \end{cases} \qquad n \geqq 1,$$

$$\int_{-\infty}^{\infty} t^{-2n+s} Q_{2n+1}(t) \, d\psi(t) = \begin{cases} 0, & 0 \leqq s \leqq 2n, \\ \gamma_{2n+1}, & s = 2n+1, \end{cases} \qquad n \geqq 0,$$

$$P_n(z) = \int_{-\infty}^{\infty} \frac{Q_n(z) - Q_n(t)}{z-t} \, d\psi(t), \qquad n \geqq 1,$$

where the constants γ_n *are defined in Theorem* 2.1.

Proof. We must look at the odd- and even-indexed polynomials separately. However, as the proofs of both cases are very similar, only the proof for odd-indexed polynomials is given here. From the results of Theorems 2.1 and 2.2 it follows that

$$\int_{-\infty}^{\infty} \frac{1}{z-t}\, d\psi(t) - L_{2n+1}(z) = \gamma_{2n+1} z^{-2n-3} + O(1/z)^{2n+4}, \qquad n \geqq 0. \tag{3.1}$$

Hence from (2.4)

$$\int_{-\infty}^{\infty} t^{2n+2}\, d\psi_{2n+1}(t) = c_{2n+2} - \gamma_{2n+1}, \qquad n \geqq 0.$$

Writing (3.1) in the form

$$\int_{-\infty}^{\infty} \frac{1}{z-t}\, d\psi(t) - L_{2n+1}(z) = z^{-2n-3} G_{2n+1}(z), \qquad n \geqq 0,$$

we have

$$G_{2n+1}(z) = \int_{-\infty}^{\infty} \frac{z t^{2n+2}}{z-t}\, d\{\psi(t) - \psi_{2n+1}(t)\}.$$

Consequently, for values of $z \in V \equiv \{z\colon z = iy,\ y > M > 0\}$,

$$|G_{2n+1}(z)| < \int_{-\infty}^{\infty} t^{2n+2}\, d\{\psi(t) + \psi_{2n+1}(t)\} = 2c_{2n+2} - \gamma_{2n+1}.$$

This indicates that the function $G_{2n+1}(z)$ is bounded, at least for values of $z \in V$. Hence in the identity

$$\int_{-\infty}^{\infty} \frac{Q_{2n+1}(z) - Q_{2n+1}(t)}{z-t}\, d\psi(t) - P_{2n+1}(z)$$

$$= z^{-2n-3} Q_{2n+1}(z) G_{2n+1}(z) - \int_{-\infty}^{\infty} \frac{Q_{2n+1}(t)}{z-t}\, d\psi(t), \qquad n \geqq 0$$

the right-hand side is a bounded function for $z \in V$ and tends to zero as $z \to \infty$ in V. But the left-hand side is a polynomial of degree less than or equal to $2n$. Therefore both sides of this identity must equal zero for all values of z to give

$$P_{2n+1}(z) = \int_{-\infty}^{\infty} \frac{Q_{2n+1}(z) - Q_{2n+1}(t)}{z-t}\, d\psi(t), \qquad n \geqq 0,$$

$$\int_{-\infty}^{\infty} \frac{Q_{2n+1}(t)}{z-t}\, d\psi(t) = z^{-2n-3} Q_{2n+1}(z) G_{2n+1}(z), \qquad n \geqq 0.$$

The first of these two equations gives the definition of $P_n(z)$ when n is odd. Expanding the second in terms of powers of $1/z$ yields

$$\begin{aligned} &\int_{-\infty}^{\infty} t Q_{2n+1}(t)\, d\psi(t) = \gamma_{2n+1}, \\ &\int_{-\infty}^{\infty} Q_{2n+1}(t)\, d\psi(t) = 0, \end{aligned} \qquad n \geqq 0. \tag{3.2}$$

Now, from Theorem 2.1 we also have

$$Q_{2n+1}(z) \int_{-\infty}^{\infty} \frac{1}{z-t}\, d\psi(t) - P_{2n+1}(z) = \delta_{2n+1} z^{2n} + O(z)^{2n+1}, \qquad n \geqq 0.$$

With the definition of $P_{2n+1}(z)$, we obtain

$$\int_{-\infty}^{\infty} \frac{Q_{2n+1}(t)}{z-t}\, d\psi(t) = \delta_{2n+1} z^{2n} + O(z)^{2n+1}, \qquad n \geqq 0.$$

Here, taking the power series expansion about the origin, we obtain

$$\int_{-\infty}^{\infty} t^{-2n+s} Q_{2n+1}(t)\, d\psi(t) = 0, \quad 0 \leqq s \leqq 2n-1, \quad n \geqq 1.$$

From this and (3.2), the definition for the odd-indexed polynomials $Q_{2n+1}(z)$, $n \geqq 0$, immediately follows. This completes the proof. □

To examine the conditions for the uniqueness of $\psi(t)$, we define the sequence of polynomials $\{Q_n(z, \tau)\}$ and $\{P_n(z, \tau)\}$ by

$$\begin{aligned} Q_{2n+1}(z, \tau) &= Q_{2n+1}(z) + \tau Q_{2n}(z), \\ P_{2n+1}(z, \tau) &= P_{2n+1}(z) + \tau P_{2n}(z), \end{aligned} \qquad n \geqq 1. \tag{3.3}$$

It is easily seen from Theorem 3.1 that these polynomials satisfy

$$\int_{-\infty}^{\infty} t^{-2n+s} Q_{2n+1}(t, \tau)\, d\psi(t) = 0, \qquad 0 \leqq s \leqq 2n-1, \tag{3.4}$$

$$P_{2n+1}(z, \tau) = \int_{-\infty}^{\infty} \frac{Q_{2n+1}(z, \tau) - Q_{2n+1}(t, \tau)}{z-t}\, d\psi(t), \tag{3.5}$$

for $n \geqq 1$, where $\psi(t)$ is any distribution with $S(\psi(t); z)$ having formal expansions $f(z)$ and $g(z)$.

It can be easily proved that, for $Q_{2n+1}(z, \tau)$, with τ taking all real values, (3.3) defines all real monic polynomials in z of degree $2n+1$ that satisfy (3.4). We may view (3.4) as a skewed quasi-orthogonality relation.

From (3.4) we find that when τ is real, all the zeros $z_r^{(2n+1)}(\tau)$, $r = 1, 2, \cdots, 2n+1$, of $Q_{2n+1}(z, \tau)$ are real and distinct. Furthermore, if $h(t)$ is any polynomial of degree less than $4n+1$, the following quadrature formula is satisfied:

$$\int_{-\infty}^{\infty} t^{-2n} h(t)\, d\psi(t) = \sum_{r=1}^{2n+1} \lambda_r^{(2n+1)}(\tau) h\{z_r^{(2n+1)}(\tau)\}, \qquad n \geqq 1,$$

where

$$\lambda_r^{(2n+1)}(\tau) = \int_{-\infty}^{\infty} t^{-2n} \left\{ \frac{Q_{2n+1}(t, \tau)}{Q'_{2n+1}\{z_r^{(2n+1)}(\tau), \tau\}\{t - z_r^{(2n+1)}(\tau)\}} \right\}^m d\psi(t) \tag{3.6}$$

for $r = 1, 2, \cdots, 2n+1$. Here m can take both values 1 and 2. Taking $m = 2$ provides the result that all $\lambda_r^{(2n+1)}(\tau)$ are positive numbers.

Consider the rational function

$$L_{2n+1}(z, \tau) = P_{2n+1}(z, \tau) / Q_{2n+1}(z, \tau),$$

and, using (3.5), expand it about infinity. We obtain

$$L_{2n+1}(z, \tau) = \frac{c_0}{z} + \frac{c_1}{z^2} + \cdots + \frac{c_{2n}}{z^{2n+1}} + O(1/z)^{2n+2}.$$

The function corresponds to the formal expansion $f(z)$. Similarly, using (3.4) and (3.5), expanding $L_{2n+1}(z, \tau)$ about the origin, we obtain

$$L_{2n+1}(z, \tau) = -c_{-1} - c_{-2} z - \cdots - c_{-2n} z^{2n-1} + O(z)^{2n},$$

provided $Q_{2n+1}(0, \tau)$ is nonzero. However, if $Q_{2n+1}(0, \tau)$ is zero, then from (2.1) we see that $Q'_{2n+1}(0, \tau)$ is nonzero, and hence this order of correspondence to the formal expansion $g(z)$ reduces by only 1.

Since the zeros of $Q_{2n+1}(z, \tau)$ are distinct, we can write

$$L_{2n+1}(z, \tau) = \sum_{r=1}^{2n+1} \frac{\rho_r^{(2n+1)}(\tau)}{z - z_r^{(2n+1)}(\tau)},$$

where $\rho_r^{(2n+1)}(\tau) = P_{2n+1}\{z_r^{(2n+1)}(\tau), \tau\}/Q'_{2n+1}\{z_r^{(2n+1)}(\tau), \tau\}$. Using (3.4)-(3.6) we can also express $\rho_r^{(2n+1)}(\tau)$ as

$$\rho_r^{(2n+1)}(\tau) = \{z_r^{(2n+1)}(\tau)\}^{2n} \lambda_r^{(2n+1)}(\tau).$$

This indicates that $\rho_r^{(2n+1)}(\tau)$ is positive unless $z_r^{(2n+1)}(\tau)$ is zero, in which case $\rho_r^{(2n+1)}(\tau)$ is also zero.

Taking the limit of $zL_{2n+1}(z, \tau)$ as $z \to \infty$, we obtain

$$\sum_{r=1}^{2n+1} \rho_r^{(2n+1)}(\tau) = c_0.$$

Hence we can write

$$L_{2n+1}(z, \tau) = \int_{-\infty}^{\infty} \frac{1}{z - t}\, d\psi_n(t, \tau),$$

where $\psi_n(t, \tau)$ is the step function

$$\psi_n(t, \tau) = \begin{cases} 0, & -\infty < t \leqq z_1^{(2n+1)}(\tau), \\ \sum_{s=1}^{r} \rho_s^{(2n+1)}(\tau), & z_r^{(2n+1)}(\tau) < t \leqq z_{r+1}^{(2n+1)}(\tau), \quad r = 1, 2, \cdots, 2n, \\ c_0, & z_{2n+1}^{(2n+1)}(\tau) < t < \infty. \end{cases}$$

These results lead to a situation similar to that of (2.4), but with rational functions of odd orders only. Since the denominator of $L_{2n+1}(z, \tau)$ with τ real represents any polynomial in z of degree $2n+1$ satisfying the skewed quasi-orthogonal relation (3.4) for any distribution whose Stieltjes function has formal expansions $f(z)$ and $g(z)$, from the above we reach the following conclusion.

The associated distribution is unique if and only if $\{L_{2n+1}(z, \tau)\}$ converges to the same limit for all real values of τ.

Since the regular real $\hat{J}$-fraction studied in [11] is equivalent to the fraction (1.4), after some simple manipulation we obtain

$$L_{2n+1}(z, \tau) = T_{2n+1}(z, w),$$

where $\tau = -\alpha_{2n+2}/w$. These functions $T_n(z, w)$ are defined on page 334 of [11] (with $a_{2n}(z) = l_{2n}$, $a_{2n+1}(z) = l_{2n+1}z$, $n \geqq 1$).

Hence, from the limit point case of $T_n(z, w)$ considered in [11], we obtain the next theorem.

THEOREM 3.2. *If the coefficients α_n and β_n satisfy, in addition to* (1.2) *and* (2.1), *one or more of the following conditions*:

I. $\sum_{r=1}^{\infty} 1/\{\alpha_{2r}\alpha_{2r+1}\}^{1/2} = \infty,$

II. $\sum_{r=1}^{\infty} |\beta_{2r}|/\{\alpha_{2r}\alpha_{2r+1}\}^{1/2} = \infty,$

III. $\beta_{2n} = 0, \quad n \geqq 1, \quad \sum_{r=1}^{\infty} \{\alpha_{2r+2}/\alpha_{2r+1}\}^{1/2} = \infty,$

then $\{L_{2n+1}(z, \tau)\}$ *converges to the same limit for all real values of* τ *(in fact, for all* τ *such that* $\operatorname{Im}(\tau) \geqq 0$*). Hence the associated distribution function* $\psi(t)$ *is unique.*

4. Special cases. First we look at a case when the distribution function $\psi(t)$ may not be unique, but does have all its points of increase in the positive half of the real axis.

To prove that the zeros of $Q_n(z)$ are real and distinct, we used the fact that $T_n(z) > 0$ for $n \geqq 1$. By including the condition $(-1)^n Q_n(0) > 0$, $n \geqq 1$, we can also easily show that the smallest zero of $Q_n(z)$ is greater than zero. Since the points of increase of $\psi_n(t)$ are the zeros of $Q_n(z)$, we have that $\psi(t)$, which is a limit of $\psi_n(t)$, also has its points of increase in the positive half of the real axis.

From (1.3) it follows that the condition $(-1)^n Q_n(0) > 0$, $n \geqq 1$, is equivalent to

$$\beta_n > 0 \quad \text{and} \quad \beta_{2n-1}\beta_{2n} - \alpha_{2n} > 0, \qquad n \geqq 1. \tag{4.1}$$

Hence we have the following theorem.

THEOREM 4.1. *If the coefficients* α_n *and* β_n *also satisfy* (4.1)*, then there exists a distribution function associated with* (1.1) *that has all its points of increase in* $(0, \infty)$.

We now consider the case in which all the coefficients α_n and β_n are bounded. If they also satisfy (1.2) and (2.1), then by Theorem 3.2 the associated distribution function is unique. We can say more about this distribution function by considering the convergence behaviour of $\{L_n(z)\}_{n=0}^{\infty}$, which are the convergents of the $\hat{J}$-fraction (1.4). We write this continued fraction in the equivalent form

$$\frac{A_1(z)}{1} + \frac{A_2(z)}{1} + \frac{A_3(z)}{1} + \frac{A_r(z)}{1} + \cdots, \tag{4.2}$$

where

$$\begin{aligned} A_1(z) &= \alpha_1/(z-\beta_1), \\ A_{2n}(z) &= -\alpha_{2n}/[\{(1+\alpha_{2n-1})z - \beta_{2n-1}\}(z-\beta_{2n})], \qquad n \geqq 1, \\ A_{2n+1}(z) &= -\alpha_{2n+1}z^2/[(z-\beta_{2n})\{(1+\alpha_{2n+1})z - \beta_{2n+1}\}], \qquad n \geqq 1. \end{aligned}$$

Hence, if there exist constants M, ε, ε_1, ε_2 of real numbers satisfying

$$0 < M < \infty, \quad 0 \leqq \varepsilon_1, \varepsilon_2 < 1, \quad 0 < \varepsilon < 1, \tag{4.3a}$$

and such that

$$\begin{aligned} &\alpha_{2n-1}/(1+\alpha_{2n-1}) \leqq (1-\varepsilon_1)(1-\varepsilon_2)(1-\varepsilon)^2, \\ &\alpha_{2n}/(1+\alpha_{2n-1}) \leqq (1-\varepsilon_1)(1-\varepsilon_2)\varepsilon^2 M^2, \\ &|\beta_{2n-1}|/(1+\alpha_{2n-1}) \leqq \varepsilon_1 M, \qquad |\beta_{2n}| \leqq \varepsilon_2 M \end{aligned} \tag{4.3b}$$

for $n \geqq 1$, then for all values of z such that $|z| \geqq M$, $A_1(z)$ is finite, $|A_{2n}(z)| \leqq \varepsilon^2$, and $|A_{2n+1}(z)| \leqq (1-\varepsilon)^2$. Here, using a well-known result (see, for example, [6, Cor. 4.36]) we obtain that for $|z| \geqq M$ all the convergents of (4.2) (i.e., (1.4)) are finite and converge also to a finite limit. We note that in the case when $\varepsilon = \frac{1}{2}$, (4.2) satisfies the Worpitzky criteria ([14, p. 42]).

This result and the fact that the zeros of $Q_n(z)$ are real and different from those of $P_n(z)$ indicate that all the zeros of $Q_n(z)$ lie in the open interval $(-M, M)$. It follows that the points of increase of $\{\psi_n(t)\}$, and hence those of its unique limit $\psi(t)$, lie in the closed interval $[-M, M]$.

Further, by using (2.3), we find that $L_n(z)$ are uniformly bounded for all values of z lying at a positive distance from the interval $[-M, M]$. Thus, from the Stieltjes-Vitali theorem, we obtain Theorem 4.2.

THEOREM 4.2. *If the coefficients α_n and β_n satisfy* (1.2), (2.1), *and* (4.3a, b) *then the associated distribution function $\psi(t)$, which is unique, has all its points of increase inside $[-M, M]$. Furthermore, the convergents $L_n(z)$ of the $\hat{J}$-fraction* (1.4) *converge uniformly to the Stieltjes function $\int_{-M}^{M} d\psi(t)/(z-t)$ in K, where K is any closed region in $\hat{\mathbb{C}}\backslash[-M, M]$. Here $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ is the extended complex plane.*

For (4.3b) to hold, it is required that α_n and β_n be bounded. On the other hand, we also can easily realise that if the coefficients α_n and β_n are bounded, then constants M, ε, ε_1, and ε_2 can be found such that (4.3) holds. Thus we get the following corollary.

COROLLARY 4.1. *If the coefficients α_n and β_n are bounded and satisfy* (1.2) *and* (2.1), *then there exists a positive number M and a unique distribution function $\psi(t)$, with all its points of increase inside $[-M, M]$, such that $\{L_n(z)\}$ converges to the Stieltjes function $\int_{-M}^{M} d\psi(t)/(z-t)$, uniformly over every finite closed region whose distance from the interval $[-M, M]$ is positive.*

We now give some examples to illustrate these results.

Example I.

$$Q_{2n}(z) = zQ_{2n-1}(z) - Q_{2n-2}(z), \qquad n \geqq 1,$$
$$Q_{2n+1}(z) = (n+1)zQ_{2n}(z) - nz^2 Q_{2n-1}(z), \qquad n \geqq 0.$$

Here $\beta_n = 0$, $\alpha_{2n} = 1$, and $\alpha_{2n+1} = n$ for $n \geqq 1$. Hence,

$$\beta_{2n-1}\beta_{2n} - \alpha_{2n} = -1, \qquad n \geqq 1,$$

and, for example,

$$\beta_{2n} = 0, \qquad n \geqq 1,$$
$$\sum_{r=1}^{\infty} \{\alpha_{2r+2}/\alpha_{2r+1}\}^{1/2} = \sum_{r=1}^{\infty} \{1/r\}^{1/2} = \infty.$$

The coefficients satisfy the conditions required by Theorem 3.2. This implies that there exists a distribution function and that it is unique.

The distribution associated with this recurrence relation is in fact $d\psi(t) = (e/\sqrt{2\pi}) \exp(-(t^2 + 1/t^2)/2)\, dt$, in $(-\infty, \infty)$ [12]. The constant $e/\sqrt{2\pi}$ is a normalising factor so that the moment $c_0 = 1$. All the moments of this distribution can be generated by the relations

$$c_0 = c_{-2} = 1, \qquad c_{2n-1} = 0, \qquad n \geqq 0,$$
$$c_{2n+2} = (2n+1)c_{2n} + c_{2n-2}, \qquad n \geqq 0,$$
$$c_{-n-2} = c_n, \qquad n \geqq 0.$$

The moments c_{2n} can also be explicitly given as

$$c_{2n} = 2^{-3n} \sum_{r=0}^{n} \left\{ \binom{2n+1}{2r+1} \sum_{s=0}^{r} \left\{ \binom{r}{s} 2^{3s} \frac{(2n-2s)!}{(n-s)!} \right\} \right\}, \qquad n \geqq 0,$$

where $\binom{p}{q}$ are the binomial coefficients.

Example II.

$$Q_{2n}(z) = (z-1)Q_{2n-1}(z) - (2n-1)Q_{2n-2}(z), \qquad n \geqq 1,$$
$$Q_{2n+1}(z) = \{(2n+1)z - (2n+2)\}Q_{2n}(z) - 2nz^2 Q_{2n-1}(z), \qquad n \geqq 0.$$

We have $\beta_{2n-1} = 2n$, $\beta_{2n} = 1$, and $\alpha_n = n - 1$, for $n \geqq 1$. Hence,

$$\beta_n > 0, \qquad \beta_{2n-1}\beta_{2n} - \alpha_{2n} = 1 > 0,$$

and also, for example,

$$\sum_{r=1}^{\infty} 1/\{\alpha_{2r}\alpha_{2r+1}\}^{1/2} = \sum_{r=1}^{\infty} 1/\{2r(2r-1)\}^{1/2} = \infty.$$

The coefficients satisfy the conditions required by Theorems 3.2 and 4.1. Hence, associated with this recurrence relation there exists a unique distribution with all its points of increase in $(0, \infty)$.

In this case we in fact have $d\psi(t) = (e/\sqrt{2\pi})t^{-1/2}\exp(-(t+1/t)/2)\,dt$ in $(0, \infty)$ (see [12]). The moments c_m associated with this distribution are the same as the even moments of the first example.

Example III. Here we take $\beta_n = 0$, $\alpha_{2n} = 1$, and

$$\alpha_{2n+1} = \lambda^2 n^2/(4n^2-1) \quad \text{for } n \geqq 1 \quad \text{where } 0 < \lambda < \infty.$$

Therefore, $\beta_{2n-1}\beta_{2n} - \alpha_{2n} = -1$, for $n \geqq 1$. Furthermore, if we let

$$b = \frac{\lambda + \sqrt{\lambda^2+4}}{2},$$

then

$$b > 1, \qquad \lambda = (b^2-1)/b,$$

$$\frac{\alpha_{2n+1}}{1+\alpha_{2n+1}} = \frac{\lambda^2 n^2}{(\lambda^2+4)n^2-1} < \frac{b^4}{(b^2+1)^2}, \qquad n \geqq 1,$$

$$\frac{\alpha_{2n+2}}{1+\alpha_{2n+1}} = \frac{4n^2-1}{(\lambda^2+4)n^2-1} < \frac{4b^2}{(b^2+1)^2}, \qquad n \geqq 1.$$

Hence, with the choice of

$$\varepsilon_1 = \varepsilon_2 = 0, \quad \varepsilon = 1/(b^2+1), \quad \text{and} \quad M = \tilde{b} = b^2+1,$$

all the conditions required by Theorem 4.2 are fulfilled. Thus we can say that the associated distribution is unique and has all its points of increase inside the interval $[-\tilde{b}, \tilde{b}]$.

The distribution associated with this recurrence relation is $d\psi(t) = dt$ in $[-b, -1/b] \cup [1/b, b]$. This distribution [12] actually has all its points of increase inside the interval $[-b, b]$. Here we can also give the relation between λ and b in the following interesting manner:

$$b = \lambda + \cfrac{1}{\lambda + \cfrac{1}{\lambda + \cfrac{1}{\lambda + \cdots}}}.$$

5. An asymptotic case. In addition to satisfying the conditions of (1.2), (2.1), and (4.3), α_n and β_n also have the following asymptotic behaviour:

$$\begin{aligned} &\lim_{n\to\infty} \alpha_{2n-1} = \alpha^{(1)}, \qquad \lim_{n\to\infty} \beta_{2n-1} = \beta^{(1)}, \\ &\lim_{n\to\infty} \alpha_{2n} = \alpha^{(2)}, \qquad \lim_{n\to\infty} \beta_{2n} = \beta^{(2)}. \end{aligned} \tag{5.1}$$

THEOREM 5.1. *Let* $Z_N \equiv \{z: z = z_r^{(n)},\ r = 1, 2, \cdots, n;\ n \geqq N\}$, *where* $z_r^{(n)}$ *are the zeros of* $Q_n(z)$. *Let* K *be any bounded closed region in* $\mathbb{C}\backslash\tilde{Z}_N$, *where* $\tilde{Z}_N$ *is the closure of* Z_N. *Then as* $n \to \infty$

$$\frac{Q_{n+2}(z)}{Q_n(z)} \to R(z) = \frac{1}{2}\left[z^2 - v_1 z - v_2 + \sqrt{\{z^2 - v_1 z - v_2\}^2 - 4v_3^2 z^2}\right], \tag{5.2}$$

uniformly in K. Here

$$v_1=\beta^{(1)}+(1+\alpha^{(1)})\beta^{(2)},\quad v_2=\alpha^{(2)}-\beta^{(1)}\beta^{(2)},\quad \textit{and}\quad v_3^2=\alpha^{(1)}\alpha^{(2)}.$$

Proof. From (1.1) we note that the even-indexed polynomials satisfy

$$Q_{2n+2}(z)=\xi_n^{(1)}(z)Q_{2n}(z)-\xi_n^{(2)}(z)Q_{2n-2}(z),\qquad n\geqq 1, \tag{5.3}$$

where

$$\xi_n^{(2)}(z)=\alpha_{2n}\alpha_{2n+1}z^2(z-\beta_{2n+2})/(z-\beta_{2n}),$$
$$\xi_n^{(1)}(z)=[(z-\beta_{2n+2})\{(1+\alpha_{2n+1})z-\beta_{2n+1}\}-\alpha_{2n+2}-\xi_n^{(2)}/\alpha_{2n}].$$

Since the coefficients α_n and β_n satisfy (1.2), (2.1), and (4.3), the zeros of $Q_{2n}(z)$ lie in the open interval $(-M, M)$ for all $n\geqq 1$. Hence for any $z\in[M,\infty)$ we arrive at the chain sequence $\{D_n(z)=d_n(z)(1-d_{n-1}(z))\}$, where

$$\xi_0^2=0,\quad D_n(z)=\xi_n^{(2)}(z)/\{\xi_n^{(1)}(z)\xi_{n-1}^{(1)}(z)\},\quad n\geqq 1,$$

with parameter sequence $\{d_n(z)\}$ given by

$$d_0(z)=0,\quad 0<d_n(z)=1-Q_{2n+2}(z)/\{\xi_n^{(1)}(z)Q_{2n}(z)\}<1,\quad n\geqq 1.$$

This parameter sequence is the minimal parameter sequence, as $d_0(z)=0$.

As α_n and β_n are asymptotically periodic and again satisfy (4.3), we find that $\{D_n(z)\}$ and hence $\{d_n(z)\}$ are convergent for any $z\in[M,\infty)$. (See Chihara [3, Thm. 6.4, p. 102] for the latter.) The convergence of $\{d_n(z)\}$ implies the convergence of $\{Q_{2n+2}(z)/Q_{2n}(z)\}$. Now, to determine the limit of $\{Q_{2n+2}(z)/Q_{2n}(z)\}$, we divide (5.3) by $Q_{2n}(z)$ and then let $n\to\infty$ obtain

$$R(z)=\{z^2-v_1z-v_2\}-v_3^2z^2/R(z). \tag{5.4}$$

Similarly, we can also obtain the exact result for the odd-indexed polynomials. Hence, for $z\in[M,\infty)$ it follows that $\{Q_{n+2}(z)/Q_n(z)\}$ converges to the limit $R(z)$ given by (5.4). Equation (5.4) gives a quadratic equation for $R(z)$, and we take (5.2) as its solution, which tends to $+\infty$ as $z\to+\infty$.

Furthermore, since $Q_n(z)$, $n\geqq N$ has no zeros in $\mathbb{C}\backslash\tilde{Z}_N$, the ratio $Q_n(z)/Q_{n+2}(z)$ is analytic in this region for $n\geqq N$. Also, since the zeros of $Q_n(z)$ are distinct and different from those of $Q_{n-1}(z)$,

$$\frac{Q_{n-1}(z)}{Q_n(z)}=\sum_{r=1}^{n}\frac{m_r^{(n)}}{z-z_r^{(n)}},\qquad n\geqq 1,$$

where

$$m_r^{(n)}=\frac{Q_{n-1}(z_r^{(n)})}{Q_n'(z_r^{(n)})}=\frac{\{Q_{n-1}(z_r^{(n)})\}^2}{T_n(z_r^{(n)})}.$$

The functions $T_n(z)$ are defined in § 2. These equations indicate that $m_r^{(n)}$ are positive and $\sum_{r=1}^{n}m_r^{(n)}=1$. Thus we can write for $z\in\mathbb{C}\backslash\tilde{Z}_N$ and $n\geqq N$

$$\left|\frac{Q_n(z)}{Q_{n+2}(z)}\right|=\left|\frac{Q_n(z)}{Q_{n+1}(z)}\right|\left|\frac{Q_{n+1}(z)}{Q_{n+2}(z)}\right|<1/\delta^2.$$

Here, δ is the minimum distance of $\tilde{Z}_N$ from z. That is to say, $Q_n(z)/Q_{n+2}(z)$ are uniformly bounded on every bounded closed region of $\mathbb{C}\backslash\tilde{Z}_N$. Therefore, by applying the Stieltjes-Vitali theorem we establish the uniform convergence of $Q_n(z)/Q_{n+2}(z)$ and hence the required results of the theorem. □

THEOREM 5.2. *In Theorem* 5.1, *if* $v_1 = 0$, *then as* $n \to \infty$

$$\frac{Q_n'(z)}{nQ_n(z)} \to \frac{R'(z)}{2R(z)} = \frac{1}{2\pi} \int_B \frac{1}{z-t} \frac{\{|t| + \mu_1\mu_2/|t|\}}{\sqrt{\mu_2^2 - t^2}\sqrt{t^2 - \mu_1^2}}\, dt$$

uniformly on every bounded closed region of $\mathbb{C}\backslash\tilde{Z}_N$, *where*

$$\mu_1 = \sqrt{v_3^2 + v_2} - v_3 \quad \text{and} \quad \mu_2 = \sqrt{v_3^2 + v_2} + v_3, \tag{5.5}$$

$$B = [-\mu_2, -\mu_1] \cup [\mu_1, \mu_2].$$

Proof. The proof of $\{Q_n'(z)/(nQ_n(z))\} \to R'(z)/(2R(z))$ for $z \in \mathbb{C}\backslash\tilde{Z}_N$, can be realised from Van Assche [13]. From (5.2),

$$\begin{aligned} R'(z) &= \tfrac{1}{2}[2z - v_1 + \tfrac{1}{2}\{(z^2 - v_1 z - v_2)^2 - 4v_3^2 z^2\}^{-1/2} \cdot \{2(z^2 - v_1 z - v_2)(2z - v_1) - 8v_3^2 z\}] \\ &= [(2z - v_1)R(z) - 2v_3^2 z]/\{(z^2 - v_1 z - v_2)^2 - 4v_3^2 z^2\}^{1/2}. \end{aligned}$$

This gives

$$\frac{R'(z)}{R(z)} = [(2z - v_1) - 2v_3^2 z/R(z)]/\{(z^2 - v_1 z - v_2)^2 - 4v_3^2 z^2\}^{1/2}.$$

However, from (5.2) and (5.4) we have

$$\frac{v_3^2 z^2}{R(z)} = \frac{1}{2}\left[z^2 - v_1 z - v_2 - \sqrt{(z^2 - v_1 z - v_2)^2 - 4v_3^2 z^2}\right].$$

Consequently

$$\frac{R'(z)}{2R(z)} = \frac{z + v_2/z}{2\sqrt{(z^2 - v_1 z - v_2)^2 - 4v_3^2 z^2}} + \frac{1}{2z}.$$

In the case where $v_1 = 0$, it follows that $v_2 > 0$ and thus

$$\frac{R'(z)}{2R(z)} = \frac{z + \mu_1\mu_2/z}{2\sqrt{(z^2 - \mu_1^2)}\sqrt{(z^2 - \mu_2^2)}} + \frac{1}{2z},$$

where μ_1 and μ_2 are defined as in (5.5). Now to complete the proof we point out the following results:

$$\frac{1}{\pi} \int_B \frac{1}{z-t} \frac{|t|}{\sqrt{\mu_2^2 - t^2}\sqrt{t^2 - \mu_1}}\, dt = \frac{z}{\sqrt{z^2 - \mu_1^2}\sqrt{z^2 - \mu_2^2}},$$

$$\frac{1}{\pi} \int_B \frac{1}{z-t} \frac{\mu_1\mu_2/|t|}{\sqrt{\mu_2^2 - t^2}\sqrt{t^2 - \mu_1^2}}\, dt = \frac{\mu_1\mu_2/z}{\sqrt{z^2 - \mu_1^2}\sqrt{z^2 - \mu_2^2}} + \frac{1}{z}.$$

The first of these results is given in Van Assche [13]. The other can be easily obtained using the first.

We now look at the partial decomposition of the ratio $Q_n'(z)/(nQ_n(z))$:

$$\frac{Q_n'(z)}{nQ_n(z)} = \sum_{r=1}^{n} \frac{1/n}{z - z_r^{(n)}} = \int_{-\infty}^{\infty} \frac{1}{z-t}\, dF_n(t).$$

The function $nF_n(t)$ can be interpreted as the number of zeros of $Q_n(z)$ less than or equal to t. Using the above theorem

$$F_n(t) \to \frac{1}{2} \int_{-\infty}^{t} \frac{|x| + \mu_1\mu_2/|x|}{\sqrt{\mu_2^2 - x^2}\sqrt{x^2 - \mu_1^2}}\, I_B\, dx,$$

where I_B is the indicator function of the set $B = [-\mu_2, -\mu_1] \cup [\mu_1, \mu_2]$.

This result indicates that if the number of zeros of $Q_n(z)$ outside B is $o(n)$ then $o(n)/n$ tends to zero as $n \to \infty$.

Example. We look at Example III in § 4. The coefficients are given by

$$\beta_n = 0, \quad \alpha_{2n} = 1, \quad \text{and} \quad \alpha_{2n+1} = \frac{\lambda^2 n^2}{4n^2 - 1}, \quad n \geqq 1.$$

Hence we have

$$\beta^{(1)} = \beta^{(2)} = 0, \quad \alpha^{(2)} = 1, \quad \text{and} \quad \alpha^{(1)} = \frac{\lambda^2}{4} = \frac{(b^2-1)^2}{4b^2}.$$

This gives

$$v_1 = 0, \quad v_2 = 1, \quad v_3 = \frac{(b^2-1)}{2b}, \quad \mu_1 = 1/b, \quad \text{and} \quad \mu_2 = b.$$

Since the associated distribution function is dt in $B = [-b, -1/b] \cup [1/b, b]$ the results are exactly as we expected.

REFERENCES

[1] O. BLUMENTHAL, *Über die Entwicklung einer willkürlichen funktion nach den Nennern des Kettenbruches für* $\int_{-\infty}^{0} [\varnothing(\xi)/(z-\xi)]\, d\xi$, Inaugural thesis, University of Göttingen, Göttingen, Germany, 1898.

[2] T. S. CHIHARA, *Orthogonal polynomials whose zeros are dense in intervals*, J. Math. Anal. Appl., 24 (1968), pp. 362-371.

[3] ———, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[4] ———, *Orthogonal polynomials with discrete spectra on the real line*, J. Approx. Theory, 42 (1984), pp. 97-105.

[5] J. FAVARD, *Sur les polynômes de Tchebicheff*, C.R. Acad. Sci. Paris, 200 (1935), pp. 2052-2053.

[6] W. B. JONES AND W. J. THRON, *Continued fractions: analytic theory and applications*, in Encyclopedia of Mathematics and Its Applications, Addison-Wesley, Reading, MA, 1980.

[7] W. B. JONES, W. J. THRON, AND H. WAADELAND, *A strong Stieltjes moment problem*, Trans. Amer. Math. Soc., 261 (1980), pp. 503-528.

[8] P. G. NEVAI, *Orthogonal Polynomials*, Mem. Amer. Math. Soc., 18 (1979).

[9] ———, *Distribution of zeros of orthogonal polynomials*, Trans. Amer. Math. Soc., 249 (1979), pp. 341-361.

[10] A. SRI RANGA, *$\hat{J}$-fractions and the strong moment problems*, in Analytic Theory of Continued Fractions II, W. J. Thron, ed., Lecture Notes in Math. 1199, Springer-Verlag, Berlin, New York, 1986, pp. 269-284.

[11] ———, *Convergence properties of a class of $\hat{J}$-fractions*, J. Comp. Appl. Math., 19 (1987), pp. 331-342.

[12] A. SRI RANGA AND J. H. MCCABE, *On the extension of some classical distributions*, Proc. Edinburgh Math. Soc. (2), to appear.

[13] W. VAN ASSCHE, *Asymptotic properties of orthogonal polynomials from their recurrence formula*, I, J. Approx. Theory, 44 (1985), pp. 258-276.

[14] H. S. WALL, *Analytic Theory of Continued Fractions*, Van Nostrand, Princeton, NJ, 1948.

SIAM J. MATH. ANAL.
Vol. 21, No. 5, pp. 1349-1368, September 1990
© 1990 Society for Industrial and Applied Mathematics
015

LINÉARISATION DE PRODUITS DE POLYNÔMES DE MEIXNER, KRAWTCHOUK, ET CHARLIER*

JIANG ZENG†

Résumé. Soit $(p_n(x))$ $(n \geq 0)$ une suite de polynômes orthogonaux par rapport à une fonctionnelle $\mathscr{L}$. Un calcul de la fonctionnelle $\mathscr{L}(\prod_{i=1}^{m} p_{n_i}(x))$ pour les polynômes de Meixner, Krawtchouk, et Charlier est proposé à l'aide de techniques combinatoires. Cette approche combinatoire permet de raffiner plusieurs résultats analytiques connus.

Abstract. Let $(p_n(x))$ $(n \geq 0)$ be a sequence of orthogonal polynomials with respect to a functional $\mathscr{L}$. A calculation of the functional $\mathscr{L}(\prod_{i=1}^{m} p_{n_i}(x))$ for the Meixner, Krawtchouk, and Charlier polynomials is proposed with the help of combinatorial techniques. This combinatorial approach permits the refinement of several classical analytical results.

Key words. linearization coefficients, orthogonal polynomials, derangements, cycles, exceedances, partitions, Pfaff-Saalschütz identity

AMS(MOS) subject classifications. 05A15, 05A17, 33A15, 33A75

1. Introduction.

Les coefficients de *linéarisation* d'une suite de polynômes orthogonaux $(p_n(x))$ sont les nombres $\alpha_{n,m,k}$ définis par: $p_n(x)p_m(x) = \sum_k \alpha_{n,m,k} p_k(x)$. De façon équivalente, si $\mathscr{L}$ est la fonctionnelle associée, le coefficient $\alpha_{n,m,k}$ est encore donné par: $\mathscr{L}(p_n p_m p_k) = \alpha_{n,m,k}\mathscr{L}(p_k p_k)$.

Beaucoup d'auteurs se sont proposés de calculer ces coefficients pour les polynômes hypergéométriques classiques en faisant apparaître des conditions simples pour la positivité de ces coefficients. Ils ont utilisé, soit des méthodes analytiques (cf. [As1], [As2], [Ea], [Ra]), soit des méthodes combinatoires (cf. [Az-Gi-Vi], [Fo-Ze2]). L'objet du présent mémoire est de reprendre à la fois les techniques analytiques et combinatoires des précédents auteurs, pour calculer effectivement $\mathscr{L}(\prod_{i=1}^{m} p_{n_i})$ pour les polynômes de Meixner, Krawtchouk, et Charlier. On obtient ainsi plusieurs formules nouvelles dans le cas où les paramètres de ces polynômes ont des valeurs arbitraires.

Dans tout cet article, on adopte la notation classique: $(a)_0 = 1$ et $(a)_n = a(a+1)\cdots(a+n-1)$, si $n \geq 1$. On suppose que m est un entier ≥ 1 fixé et que $\mathbf{n} = (n_1, \cdots, n_m)$ est une suite de m entiers positifs. On note $\mathbf{n}^* = (n_{i_1} \geq n_{i_2} \geq \cdots \geq n_{i_m})$ le réarrangement décroissant de $\mathbf{n}$. Enfin, $[n]$ désigne l'intervalle $\{1, 2, \cdots, n\}$ des entiers.

Rappelons que les polynômes de Meixner $M_n(x; \beta, c)$ sont définis par

$$M_n(x; \beta, c) = \sum_{k=0}^{n} \binom{n}{k} (-x)_k (\beta + k)_{n-k} (c^{-1} - 1)^k \qquad (n \geq 0) \tag{1.1}$$

(cf. [Ch]), où $\beta > 0$ et $0 < c < 1$. Ils sont orthogonaux par rapport à la fonction-poids discrète $(c^k(\beta)_k/k!)$ $(k \geq 0)$, portée par $\mathbf{N}$.

Les polynômes de Krawtchouk $(K_n(x; p, N))$ sont définis par:

$$K_n(x; p, N) = \sum_{k=0}^{N} \frac{(-n)_k (-x)_k}{(-N)_k k!} \left(\frac{1}{p}\right)^k, \tag{1.2}$$

* Received by the editors June 29, 1988; accepted for publication (in revised form) November 9, 1989.
† Département de mathématique, Université Louis-Pasteur, 7 rue René-Descartes, F-67084 Strasbourg, France. This work was supported by the French Coordinated Research Program in Mathematics and Computer Science.

où $0<p<1$ et $n=0,1,\cdots,N$. Ils sont orthogonaux par rapport à la fonction-poids $\binom{N}{x}p^x(1-p)^{N-x}$, portée par les entiers $0,1,\cdots,N$ (cf. [Ch]).

Les polynômes de Charlier $C_n^{(a)}(x)$ sont eux définis par:

$$C_n^{(a)}(x)=\sum_{k=0}^{n}\binom{n}{k}(-x)_k a^{-k} \qquad (a>0,\ n\geqq 0) \tag{1.3}$$

et sont orthogonaux par rapport à la fonction-poids $a^k/k!$ portée par $\mathbf{N}$ (cf. [Ch]).

Les trois fonctionnelles correspondant à ces trois suites de polynômes sont données par:[1]

$$\mathcal{M}(\mathbf{n};\beta,c)=(-1)^{n_1+\cdots+n_m}(1-c)^{\beta}\sum_{x\geqq 0}\prod_{i=1}^{m}M_{n_i}(x;\beta,c)\frac{c^x(\beta)_x}{x!}, \tag{1.4}$$

$$\mathcal{K}(\mathbf{n};N,p)=\prod_{i=1}^{m}(-1)^{n_i}(-N)_{n_i}\sum_{x=0}^{N}\prod_{i=1}^{m}K_{n_i}(x;p,N)\binom{N}{x}p^x(1-p)^{N-x}, \tag{1.5}$$

$$\mathcal{C}(\mathbf{n};a)=e^{-a}\prod_{i=1}^{m}(-a)^{n_i}\sum_{x\geqq 0}\prod_{i=1}^{m}C_{n_i}^{(a)}(x)\frac{a^x}{x!}. \tag{1.6}$$

Nos résultats principaux concernant les fonctionnelles de Meixner et Krawtchouk reposent sur des propriétés statistiques d'une classe d'objets combinatoires, appelés *dérangements colorés*. On les introduit de la façon suivante. D'abord toute suite $\mathbf{n}=(n_1,\cdots,n_m)$ de m entiers positifs de somme n détermine, de façon unique, une application χ de $[n]$ sur $[m]$, donnée par $\chi(j)=i$, si $n_1+\cdots+n_{i-1}<j\leqq n_1+\cdots+n_i$ $(1\leqq i\leqq m$; par convention, $n_0=0)$. On dit que χ est le *m-coloriage* de $[n]$ associé à la suite $\mathbf{n}$. On note $C_{\mathbf{n}}$ l'ensemble de tous les couples $(\chi(j),j)$, $(j=1,\cdots,n)$. Par commodité, $\chi(j)$ est appelé la *couleur* de j. Soit π une permutation de $C_{\mathbf{n}}$; par abus de notation, on pose:

$$\pi(\chi(j),j)=(\chi(\pi(j)),\pi(j));$$

son *nombre de cycles* est noté $\operatorname{cyc}\pi$; on dit que π est un *dérangement* (*coloré*) de $C_{\mathbf{n}}$ si $\chi(j)\neq\chi(\pi(j))$ pour tout $j\in[n]$; on note $\mathcal{D}(\mathbf{n})$ l'ensemble des dérangements de $C_{\mathbf{n}}$. On dit, d'autre part, que π a une *excédance* en j $(1\leqq j\leqq n)$, si $\chi(j)<\chi(\pi(j))$; on note $\operatorname{exc}\pi$ le nombre des excédances de π. On définit enfin le *poids* de π par: $w(\pi)=\beta^{\operatorname{cyc}\pi}\gamma^{\operatorname{exc}\pi}$. Le polynôme générateur de $\mathcal{D}(\mathbf{n})$ par w est alors défini par:

$$D(\mathbf{n};\beta,\gamma)=\sum_{\pi}w(\pi) \qquad (\pi\in\mathcal{D}(\mathbf{n})).$$

Pour l'étude de la fonctionnelle de Charlier, on utilise la notion de partition de l'ensemble $C_{\mathbf{n}}$. Une telle partition π est dite *partition colorée*, si chaque bloc de π est constitué seulement par des éléments de couleurs différentes. Le nombre de blocs dans la partition π est noté $\operatorname{bloc}\pi$ et $\mathcal{PCS}(\mathbf{n})$ désigne l'ensemble des partitions colorées sans points isolés (c'est-à-dire, sans blocs réduits à une élément) de $C_{\mathbf{n}}$. Chaque partition π de $C_{\mathbf{n}}$ est munie d'un poids défini par: $\nu(\pi)=a^{\operatorname{bloc}\pi}$. Le polynôme générateur de $\mathcal{PCS}(\mathbf{n})$ par ν est défini par:

$$PCS(\mathbf{n};a)=\sum_{\pi}\nu(\pi) \qquad (\pi\in\mathcal{PCS}(\mathbf{n})).$$

Le résultat essentiel de notre article consiste à établir les trois identités:

$$\mathcal{M}(\mathbf{n};\beta,c)=D(\mathbf{n};\beta,c^{-1}), \tag{1.7}$$

[1] Les fonctionnelles de Meixner et de Krawtchouk auront le même support combinatoire. Pour des raisons d'homogénéité, on a dû adopter la notation $\mathcal{K}(\mathbf{n};N,p)$ pour la fonctionnelle de Krawtchouk, alors que le polynôme est noté classiquement $K_n(x;p,N)$.

$$\mathcal{K}(\mathbf{n};\, N,\, p) = D(\mathbf{n};\, -N,\, 1-1/p), \tag{1.8}$$

$$\mathcal{C}(\mathbf{n};\, a) = PCS(\mathbf{n};\, a). \tag{1.9}$$

La première identité montre que $\mathcal{M}(\mathbf{n};\, \beta,\, c)$ est un polynôme de variables β et c^{-1} à coefficients entiers positifs; donc la positivité de $\mathcal{M}(\mathbf{n};\, \beta,\, c)$ lorsque $\beta \geqq 0$ et $0 < c < 1$ est évidente: $\mathcal{M}(\mathbf{n};\, \beta,\, c) \geqq 0$.

La seconde identité indique que $\mathcal{K}(\mathbf{n};\, N,\, p)$ est un polynôme en les variables $(-N)$ et $(1-1/p)$ à coefficients entiers positifs, mais cette fois-ci l'interprétation de $\mathcal{K}(\mathbf{n};\, N,\, p)$ ne donne pas d'information sur son signe immédiatement. On a besoin de faire appel à l'algèbre des fonctions symétriques.

Comme démontré dans le Corollaire 10, on a la forme explicite de la fonctionnelle sous la forme:

$$\mathcal{K}(\mathbf{n};\, N,\, p) = \prod_{i=1}^{m} n_i! \sum_{s \geqq 0} (-N)_s (p-1)^s \sum_{\lambda} a_{\lambda\mu} \prod_{i=2}^{m} \frac{(1-(1-1/p)^{i-1})^{k_i}}{k_i!}, \tag{1.10}$$

où $\mu = \mathbf{n}^*$ et λ varie dans l'ensemble des partitions $(1^0 2^{k_2} \cdots m^{k_m})$ de l'entier $n = n_1 + \cdots + n_m$ satisfaisant $k_2 + \cdots + k_m = s$ et où les $a_{\lambda\mu}$ sont des coefficients de changement de base s'exprimant à l'aide des nombres de Kostka (cf. Corollaire 10). On montrera comment cette formule permet d'établir directement un résultat fondamental sur la positivité de la fonctionnelle de Krawtchouk (cf. Corollaire 11).

La troisième identité montre que $\mathcal{C}(\mathbf{n};\, a)$ est un polynôme en la variable a, à coefficients entiers positifs. La fonctionnelle correspondante est donc positive lorsque $a \geqq 0$.

Les techniques de démonstration font appel aux méthodes du composé partitionnel (cf. [Fo-Sch], [Fo]) et d'autre part prolongent, dans un contexte combinatoire, des calculs analytiques faits par Askey et Ismail [As-Is]. Ces deux derniers auteurs ont établi l'identité (1.7) lorsque $\beta = 1$. La formule (1.10), valable pour m quelconque se spécialise pour $m = 3$ en une formule obtenue par Askey et Gasper [As-Ga]. L'interprétation combinatoire des fonctionnelles de Krawtchouk et Charlier est nouvelle.

Après la présente introduction, l'article s'ouvre sur trois sections et un appendice. La deuxième section contient toutes les techniques combinatoires utilisées. On y prolonge un calcul de déterminant introduit par Askey et Ismail [As-Is] (cf. Théorème 1), à l'aide de la β-extension du "Master Theorem," établi par Foata et Zeilberger [Fo-Ze2]. En fait, pour mener le calcul de la fonctionnelle à son terme, il faut non seulement évaluer le polynôme générateur des dérangements, mais trouver la relation entre ce dernier et le polynôme générateur de toutes les permutations (voir Théorèmes 3 et 4).

Il suffit, dans § 3, d'appliquer les résultats établis dans la section précédente pour démontrer les trois identités (1.7), (1.8), et (1.9).

Il est remarquable de constater que *toutes* les fonctionnelles des polynômes orthogonaux hypergéométriques classiques sont des fonctions génératrices de *dérangements*. La section 4 montre qu'il y a une cohérence totale entre les diverses interprétations combinatoires de ces fonctionnelles. On retrouve, pour celles-ci, le même tableau qu'Askey et Wilson [As-Wi] avaient construit pour les polynômes eux-mêmes.

L'appendice propose une démonstration analytique d'une identité (Corollaire 2), déjà établie à l'aide de techniques combinatoires.

2. Calcul permutationnel et partitionnel.

2.1. Trois lemmes fondamentaux. Soient A un ensemble fini et S un sous-ensemble de A. Chaque injection π de S dans A peut s'identifier avec son graphe: les sommets

sont les éléments de A, les arcs sont les flèches allant de i à j si $\pi(i)=j$. On désigne par cyc π le nombre de cycles du graphe et l'on définit le poids de π par $w(\pi)=\beta^{\mathrm{cyc}\,\pi}$. On note enfin Inj (S, A) l'ensemble des injections de S dans A et $|A|$ le cardinal de A. Le polynôme générateur de Inj (S, A) par le nombre de cycles prend la forme simple exprimée dans le lemme suivant (cf. [Fo-St]).

LEMME 1. *Soient* $0\leqq k\leqq n$, $|A|=n$ *et* $|S|=n-k$, *on a*

$$w(\mathrm{Inj}\,(S,A))=\sum_{\pi} w(\pi)=(\beta+k)_{n-k} \qquad (\pi\in \mathrm{Inj}\,(S,A)). \tag{2.1}$$

Remarque. Lorsque $k=0$, à savoir $S=A$, l'ensemble Inj (S, A) se compose des permutations de A. La formule (2.1) se réduit alors au résultat classique (cf. [Ri]):

$$\sum_{\pi} w(\pi)=\sum_{\pi}\beta^{\mathrm{cyc}\,\pi}=(\beta)_n \qquad (\pi\in S_n). \tag{2.2}$$

Dans l'introduction, on a défini l'ensemble $C_{\mathbf{n}}$ de tous les couples $(\chi(j), j)$, $(j=1,\cdots,n)$, ainsi que l'ensemble des *dérangements* $\mathscr{D}(\mathbf{n})$ de $C_{\mathbf{n}}$, définis à l'aide de la fonction χ. On introduit, en plus, l'ensemble $\mathscr{P}(\mathbf{n})$ des permutations de $C_{\mathbf{n}}$. On munit ensuite chaque permutation π de $C_{\mathbf{n}}$ du poids $\nu(\pi)$ défini par

$$\nu(\pi)=\beta^{\mathrm{cyc}\,\pi}\prod_{j=1}^{n} b(\chi(j),\chi(\pi(j))) \tag{2.3}$$

et on pose:

$$\nu(\mathscr{P}(\mathbf{n}))=\sum_{\pi}\nu(\pi) \qquad (\pi\in\mathscr{P}(\mathbf{n})). \tag{2.4}$$

La fonction génératrice de $\nu(\mathscr{P}(\mathbf{n}))$ a une forme explicite donnée par la β-extension du "Master Theorem" (cf. [Fo-Ze]). Dans l'énoncé suivant, V_m désigne le déterminant classique du "Master Theorem," à savoir, $\det(\delta_{ij}-b(i,j)x_j)$ $(1\leqq i, j\leqq m)$.

LEMME 2 (β-extension du "Master Theorem"). *On a l'identité*:

$$\sum\frac{x_1^{n_1}}{n_1!}\cdots\frac{x_m^{n_m}}{n_m!}\nu(\mathscr{P}(\mathbf{n}))=V_m^{-\beta}, \tag{2.5}$$

où la sommation de gauche est faite sur toutes les suites $\mathbf{n}=(n_1,\cdots,n_m)$ *de m entiers positifs.*

Remarque. Prolongeant le résultat du Lemme 2, nous avons, par ailleurs (cf. [Ze2]), donné une β-extension de la formule d'inversion de Lagrange à plusieurs variables.

La formule exponentielle va jouer un rôle important dans cet article. Nous rappelons ici la méthode du *composé partitionnel* (cf. [Fo-Sch], [Fo2]) permettant un calcul simple de plusieurs fonctions génératrices exponentielles. Étant donné un ensemble fini A, une *partition* de A est une collection $\pi=\{S_1,\cdots,S_r\}$ de ses sous-ensembles non vides, mutuellement disjoints, dont l'union est égale à A. On appelle *bloc* chaque part S_i dans π, et l'on note bloc π le nombre de blocs de π. On note enfin $\Pi[n]$ (respectivement, $S[n]$) l'ensemble des partitions (respectivement, partitions en un bloc) de $[n]$ et on pose

$$\Pi=\bigcup_{n\geqq 1}\Pi[n], \qquad S=\bigcup_{n\geqq 1}S[n].$$

On peut identifier Π au *composé partitionnel abélien* $S^{(+)}$ de S (cf. [Fo2]).

Rappelons qu'une *application multiplicative* est une application μ de $S^{(+)}$ dans une algèbre de polynômes, telle que si π est une partition $\{S_1,\cdots,S_r\}$, où chaque S_i

est un bloc de taille n_i, le polynôme $\mu(\pi)$ est donné par le produit $\mu(S[n_1])\cdots\mu(S[n_r])$. On peut alors considérer les polynômes:

$$\mu\{\Pi[n]\}=\sum\{\mu(\pi):\ \pi\in\Pi[n]\}\quad\text{et}\quad\mu\{S[n]\}=\sum\{\mu(\pi):\ \pi\in S[n]\}.$$

Comme démontré dans [Fo2], ils sont reliés par la formule exponentielle exprimée ci-après.

LEMME 3. *Si μ est une application multiplicative, on a l'identité*:

$$1+\sum_{n\geqq 1}\mu\{\Pi(n]\}\frac{u^n}{n!}=\exp\left(\sum_{n\geqq 1}\mu\{S[n]\}\frac{u^n}{n!}\right). \tag{2.6}$$

2.2. Permutations et dérangements. Dans l'introduction on a défini le poids

$$w(\pi)=\beta^{\operatorname{cyc}\pi}\gamma^{\operatorname{exc}\pi} \tag{2.7}$$

d'une permutation π de $C_{\mathbf{n}}$. On a aussi défini le polynôme générateur $\mathscr{D}(\mathbf{n})$. Nous rappelons ci-après cette définition et introduisons, en plus, le polynôme générateur $P(\mathbf{n};\beta,\gamma)$ de *toutes* les permutations de $C_{\mathbf{n}}$:

$$\begin{aligned}P(\mathbf{n};\beta,\gamma)&=\sum_{\pi}w(\pi)\qquad(\pi\in\mathscr{P}(\mathbf{n})),\\ \mathrm{D}(\mathbf{n};\beta,\gamma)&=\sum_{\pi}w(\pi)\qquad(\pi\in\mathscr{D}(\mathbf{n})).\end{aligned} \tag{2.8}$$

Les fonctions génératrices de ces deux polynômes ont des formes explicites, qui sont exprimées ci-dessous à l'aide des fonctions symétriques élémentaires $e_1,\cdots,e_m$ des variables $x_1,\cdots,x_m$.

THÉORÈME 1. *On a les identités*

$$\begin{aligned}&\sum_{\mathbf{n}}\frac{x_1^{n_1}}{n_1!}\cdots\frac{x_m^{n_m}}{n_m!}P(\mathbf{n};\beta,\gamma)\\ &\quad=[1-e_1-(\gamma-1)e_2-\cdots-(\gamma-1)^{m-1}e_m]^{-\beta},\end{aligned} \tag{2.9}$$

$$\begin{aligned}&\sum_{\mathbf{n}}\frac{x_1^{n_1}}{n_1!}\cdots\frac{x_m^{n_m}}{n_m!}D(\mathbf{n};\beta,\gamma)\\ &\quad=[1-\gamma e_2-\gamma(1+\gamma)e_3-\cdots-\gamma(1+\gamma+\cdots+\gamma^{m-2})e_m]^{-\beta},\end{aligned} \tag{2.10}$$

où les sommations sont faites sur toutes les suites $\mathbf{n}=(n_1,\cdots,n_m)$ de m entiers positifs.

Remarque. Lorsque $\beta=1$, Askey et Ismail [As-Is] ont obtenu la seconde identité avec une interprétation combinatoire analogue.

Démonstration. Pour $1\leqq i,j\leqq m$, posons

$$b(i,j)=\begin{cases}\gamma & \text{si } i<j,\\ 1 & \text{sinon.}\end{cases} \tag{2.11}$$

En substituant $b(i,j)$ dans le Lemme 2 et en posant $n=n_1+\cdots+n_m$, on trouve:

$$\begin{aligned}\nu(\mathscr{P}(\mathbf{n}))&=\sum_{\pi\in\mathscr{P}(\mathbf{n})}\beta^{\operatorname{cyc}\pi}\prod_{j=1}^{n}b(\chi(j),\chi(\pi(j)))\\ &=\sum_{\pi\in\mathscr{P}(\mathbf{n})}\beta^{\operatorname{cyc}\pi}\gamma^{\operatorname{exc}\pi}.\end{aligned}$$

Le déterminant $\det(\delta_{i,j}-b(i,j)x_j)$ devient:

$$V_m=(-1)^m x_1\cdots x_m\begin{vmatrix} 1-x_1^{-1} & \gamma & \cdots & \gamma & \gamma \\ 1 & 1-x_2^{-1} & \cdots & \gamma & \gamma \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 1-x_{m-1}^{-1} & \gamma \\ 1 & 1 & \cdots & 1 & 1-x_m^{-1} \end{vmatrix}$$

$$=1-e_1-(\gamma-1)e_2-\cdots-(\gamma-1)^{m-1}e_m$$

(cf., par exemple, Muir [Mu, p. 441]).

L'identité (2.10) est démontrée de façon analogue. Il suffit de prendre pour $b(i,j)$ les valeurs:

$$b(i,j)=\begin{cases}\gamma & \text{si } i<j,\\ 0 & \text{si } i=j,\\ 1 & \text{si } i>j.\end{cases}$$

On notera qu'à cause des valeurs $b(i,i)=0$, la somme $\nu(\mathscr{P}(\mathbf{n}))$ est égale à $D(\mathbf{n};\beta,\gamma)$. Le déterminant de MacMahon devient:

$$V_m=(-1)^m x_1\cdots x_m\begin{vmatrix} -x_1^{-1} & \gamma & \cdots & \gamma & \gamma \\ 1 & -x_2^{-1} & \cdots & \gamma & \gamma \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & -x_{m-1}^{-1} & \gamma \\ 1 & 1 & \cdots & 1 & -x_m^{-1} \end{vmatrix}.$$

qu'on peut encore exprimer sous la forme:

$$V_m=1-\gamma e_2-\gamma(1+\gamma)e_3-\cdots-\gamma(1+\gamma+\cdots+\gamma^{m-2})e_m$$

(cf. Muir [Mu, p. 441] et Askey et Ismail [As-Is, § 3]). □

L'identité (2.9) permet de donner une formule explicite pour le polynôme $P(\mathbf{n};\beta,\gamma)$ donnée dans le corollaire suivant.

COROLLAIRE 1. *On a l'identité*

$$P(\mathbf{n};\beta,\gamma)=(1-\gamma^{-1})^{\beta}(1-\gamma)^n\sum_{k\geqq 0}\prod_{i=1}^{m}(-k)_{n_i}\frac{\gamma^{-k}(\beta)_k}{k!}, \tag{2.12}$$

où $n=n_1+\cdots+n_m$.

Remarque. Lorsque $m=1$, en appliquant la formule binomiale, on retrouve (2.2). Lorsque tous les n_i valent 1, le premier membre de (2.12) devient le polynôme générateur des permutations de $[m]$ par le nombre des cycles et celui des excédances. Or, d'après la transformation fondamentale de Foata [Fo2], ce dernier est aussi le polynôme générateur des permutations par le nombre des éléments *saillants* et celui des *descentes*, on retrouve donc une formule apparaissant dans le mémoire de Viennot [Vi, formule (66), p. II-37].

Démonstration. Il suffit de vérifier que le membre de droite de (2.12) satisfait la formule (2.9). En effet, en substituant $P(\mathbf{n};\beta,\gamma)$ dans (2.9) par le membre de droite

de (2.12), on a

$$(1-\gamma^{-1})^{\beta}\sum_{k\geqq 0}\left(\prod_{i=1}^{m}\sum_{n_i\geqq 0}(1-\gamma)^{n_i}(-k)_{n_i}\frac{x_i^{n_i}}{n_i!}\right)\frac{\gamma^{-k}(\beta)_k}{k!}$$

$$=(1-\gamma^{-1})^{\beta}\sum_{k\geqq 0}\prod_{i=1}^{m}(1-(1-\gamma)x_i)^{k}\frac{\gamma^{-k}(\beta)_k}{k!}$$

$$=(1-\gamma^{-1})^{\beta}\left[\left(1-\gamma^{-1}\prod_{i=1}^{m}(1-(1-\gamma)x_i)\right)\right]^{-\beta}$$

$$=[1-e_1-(\gamma-1)e_2-\cdots-(\gamma-1)^{m-1}e_m]^{-\beta},$$

en appliquant par deux fois la formule binomiale. □

Dans (2.12), en faisant γ tendre vers 1 et en utilisant l'identité (2.2), on obtient le corollaire ci-dessous, qui permet d'interpréter la fonctionnelle de produits de polynômes de Laguerre (voir Théorème 8 ci-après) à partir de celle de Meixner.

COROLLAIRE 2. *On a*

$$\lim_{\gamma\to 1}(1-\gamma^{-1})^{\beta}(1-\gamma)^{n_1+\cdots+n_m}\sum_{k\geqq 0}\prod_{i=1}^{m}(-k)_{n_i}\frac{\gamma^{-k}(\beta)_k}{k!}=(\beta)_{n_1+\cdots+n_m}.$$

On se propose maintenant de trouver une formule polynomiale à coefficients positifs pour $D(\mathbf{n};\beta,\gamma)$ dans le cas où m est quelconque. Pour cela on a besoin d'un résultat sur les fonctions symétriques. Soient

$$\lambda=(\lambda_1\geqq\lambda_2\geqq\cdots)\quad\text{et}\quad\mu=(\mu_1\geqq\mu_2\geqq\cdots)$$

deux *partitions* d'entiers, la plus grande part étant $\leqq m$. Comme il est bien connu (cf. [Macd, p. 65]), chaque élément $e_\lambda=e_{\lambda_1}e_{\lambda_2}\cdots$ s'exprime à l'aide des fonctions symétriques monomiales m_μ par l'équation

$$e_\lambda=\sum_{\mu}a_{\lambda\mu}m_\mu, \tag{2.13}$$

où pour tout couple de partitions (λ,μ), le coefficient $a_{\lambda\mu}$ est le nombre des matrices $(0,1)$ dont les sommes de ligne sont les λ_i et les sommes de colonnes les μ_j. Les coefficients $a_{\lambda\mu}$ sont donc des entiers positifs, qui ont une expression simple en fonction des nombres de Kostka (cf. [Macd, p. 65]).

COROLLAIRE 3. *On a l'identité*

$$D(\mathbf{n};\beta,\gamma)=\left(\prod_{i=1}^{m}n_i!\right)\sum_{s\geqq 0}(\beta)_s\left(\frac{\gamma}{1-\gamma}\right)^{s}\sum_{\lambda}a_{\lambda\mu}\prod_{i=2}^{m}\frac{(1-\gamma^{i-1})^{k_i}}{k_i!}, \tag{2.14}$$

où $\mu=\mathbf{n}^*$ et où la seconde sommation est faite sur toutes les partitions $\lambda=(1^0 2^{k_2}\cdots m^{k_m})$ de l'entier $n_1+\cdots+n_m$ satisfaisant $k_2+\cdots+k_m=s$.

Démonstration. On développe le membre de droite de (2.10) à l'aide de la formule binomiale, puis à l'aide de la formule multinomiale. On obtient:

$$[1-\gamma e_2-\gamma(1+\gamma)e_3-\cdots-\gamma(1+\gamma+\cdots+\gamma^{m-2})e_m]^{-\beta}$$

$$=\sum_{s}\frac{(\beta)_s}{s!}\frac{\gamma^s}{(1-\gamma)^2}\sum\frac{s!}{k_2!k_3!\cdots k_m!}\prod_{i=2}^{m}(1-\gamma^{i-1})^{k_i}e_\lambda,$$

où $k_2+\cdots+k_m=s$ et $\lambda=(1^0 2^{k_2}\cdots m^{k_m})$. On exprime ensuite les e_λ en termes des fonctions symétriques monomiales m_μ et l'on l'identifie avec le premier membre de l'identité (2.10). □

Remarque. Lorsque $m=3$, la formule (2.14) se spécialise en une formule apparaissant comme une *β-extension* de l'identité de Pfaff-Saalschütz (voir § 2.5, Proposition 3). On peut aussi obtenir une formule analogue pour $P(\mathbf{n}; \beta, \gamma)$.

2.3. Partitions colorées. Considérons de nouveau l'ensemble $C_{\mathbf{n}}$ associé à la suite $\mathbf{n}=(n_1, \cdots, n_m)$ de m entiers positifs. Comme déjà dit dans l'introduction, une partition π de $C_{\mathbf{n}}$ est dite *partition colorée*, si chaque bloc de π est constitué seulement par des éléments de couleurs différentes. Notons $\mathscr{PC}(\mathbf{n})$ (respectivement, $\mathscr{PCS}(\mathbf{n})$) l'ensemble des partitions colorées (respectivement, partitions colorées sans points isolés) de $C_{\mathbf{n}}$ et introduisons pour chaque partition colorée π de $C_{\mathbf{n}}$ le poids:

$$\nu(\pi)=a^{\text{bloc}\,\pi}.$$

Les polynômes générateurs de $\mathscr{PC}(\mathbf{n})$ et $\mathscr{PCS}(\mathbf{n})$ sont définis respectivement par:

$$PC(\mathbf{n}; a)=\sum_{\pi} \nu(\pi) \qquad (\pi \in \mathscr{PC}(\mathbf{n})),$$

$$PCS(\mathbf{n}; a)=\sum_{\pi} \nu(\pi) \qquad (\pi \in \mathscr{PCS}(\mathbf{n})).$$

A l'aide du Lemme 3, on peut donner des formules explicites pour les fonctions génératrices de ces deux polynômes.

THÉORÈME 2. *On a les identités*:

$$\sum_{\mathbf{n}} \frac{x_1^{n_1}}{n_1!} \cdots \frac{x_m^{n_m}}{n_m!} PC(\mathbf{n}; a)=\exp(ae_1+\cdots+ae_m), \tag{2.15}$$

$$\sum_{\mathbf{n}} \frac{x_1^{n_1}}{n_1!} \cdots \frac{x_m^{n_m}}{n_m!} PCS(\mathbf{n}; a)=\exp(ae_2+\cdots+ae_m), \tag{2.16}$$

où les sommations sont faites sur toutes les suites $\mathbf{n}=(n_1, \cdots, n_m)$ *de m entiers positifs.*

On présente ici deux démonstrations, l'une s'appuie sur la formule exponentielle à une variable, et l'autre est de nature plus combinatoire et élémentaire.

Démonstration 1. Pour chaque partition π de $[n]$ et chaque application φ de $[n]$ dans $[m]$, on définit:

$$\theta(\pi, \varphi)=\begin{cases}1 & \text{si la restriction de } \varphi \text{ à chaque bloc de } \pi \text{ est injective,}\\ 0 & \text{sinon.}\end{cases}$$

L'application θ permet d'introduire pour chaque $\pi \in \Pi[n]$ le poids comme suit:

$$\mu(\pi)=\sum_{\varphi:[n]\to[m]} \theta(\pi, \varphi) a^{\text{bloc}\,\pi} \prod_{i=1}^{n} x_{\varphi(i)}.$$

On vérifie facilement que μ est une application multiplicative. Ceci revient à dire que $\mu(\pi)$ est le produit de tous les poids des blocs. Par ailleurs, pour toute suite $\mathbf{n}=(n_1, \cdots, n_m)$ de m entiers positifs de somme n, posons

$$\nu\{\Pi[n]\}=\sum_{\pi} \theta(\pi, \chi) a^{\text{bloc}\,\pi} \qquad (\pi \in \Pi[n]),$$

$$\nu\{S[n]\}=\sum_{\pi} \theta(\pi, \chi) a^{\text{bloc}\,\pi} \qquad (\pi \in S[n]),$$

où χ est le m-coloriage de $[n]$ associé à la suite $\mathbf{n}$. On vérifie également:

$$\mu\{\Pi[n]\}=\sum_{\mathbf{n}} \frac{n!}{n_1! \cdots n_m!} x_1^{n_1} \cdots x_1^{n_m} \nu\{\Pi[n]\}, \tag{2.17}$$

$$\mu\{S[n]\} = \sum_{\mathbf{n}} \frac{n!}{n_1! \cdots n_m!} x_1^{n_1} \cdots x_m^{n_m} \nu\{S[n]\}, \tag{2.18}$$

où les sommations de droite sont faites sur toutes les suites $\mathbf{n} = (n_1, \cdots, n_m)$ de m entiers positifs de somme n. En substituant (2.17) et (2.18) dans (2.6) et en faisant $u = 1$, on obtient

$$\sum_{\mathbf{n}} \frac{x_1^{n_1}}{n_1!} \cdots \frac{x_m^{n_m}}{n_m!} \nu\{\Pi[n_1 + \cdots + n_m]\} = \exp \sum_{\mathbf{n}} \frac{x_1^{n_1}}{n_1!} \cdots \frac{x_m^{n_m}}{n_m!} \nu\{S[n_1 + \cdots + n_m]\},$$

où la sommation de gauche (respectivement, droite) est faite sur toutes les suites $\mathbf{n} = (n_1, \cdots, n_m)$ de m entiers positifs (respectivement, de somme $\geqq 1$). On remarque d'abord: $\nu\{\Pi[n_1 + \cdots + n_m]\} = PC(\mathbf{n}; a)$; d'autre part, $\nu\{S[n_1 + \cdots + n_m]\} = 0$ sauf si tous les n_i valent 0 ou 1, il en résulte: $\nu\{S[n_1 + \cdots + n_m]\} = a\sigma_1 + \cdots + a\sigma_m$. Ce qui établit (2.15).

On démontre l'identité (2.16) de façon analogue. Il suffit de prendre pour θ les valeurs suivantes:

$$\theta(\pi, \varphi) = \begin{cases} 1 & \text{si } \pi \text{ n'a pas de point isolé} \\ & \text{et si la restriction de } \varphi \text{ à chaque bloc de } \pi \text{ est injective,} \\ 0 & \text{dans le cas contraire.} \end{cases}$$

□

Démonstration 2. En identifiant les coefficients de $x_1^{n_1} \cdots x_m^{n_m}$ des deux membres de (2.15), on constate que (2.15) est équivalente à

$$PC(\mathbf{n}; a) = \sum_{s \geqq 0} a^s \sum_{\lambda} \left(\prod_{i=1}^{m} \frac{n_i!}{k_i!} \right) a_{\lambda\mu}, \tag{2.15'}$$

où $\lambda = (1^{k_1} 2^{k_2} \cdots m^{k_m})$ avec $k_1 + \cdots + k_m = s$ et $\mu = \mathbf{n}^*$ et enfin où $a_{\lambda\mu}$ est défini par $e_\lambda = \sum a_{\lambda,\mu} m_\mu$. Donc pour prouver (2.15) il suffit de démontrer que le nombre des partitions colorées de $C_{\mathbf{n}}$ à s blocs dont k_i blocs contiennent exactement i éléments $(1 \leqq i \leqq m)$ est égal à

$$\left(\prod_{i=1}^{m} \frac{n_i!}{k_i!} \right) a_{\lambda\mu}.$$

Ceci est évident car $a_{\lambda\mu}$ est le nombre de matrices (0, 1) dont les sommes de ligne sont les λ_i $(1 \leqq i \leqq s)$ et les sommes de colonnes les μ_j $(1 \leqq j \leqq m)$. La formule (2.16) peut s'établir de la même façon. □

A l'aide du Théorème 2, on peut donner une formule explicite du polynôme générateur $PC(\mathbf{n}; a)$.

COROLLAIRE 4. *On a l'identité*:

$$PC(\mathbf{n}; a) = (-1)^{n_1 + \cdots + n_m} e^{-a} \sum_{k \geqq 0} \prod_{i=1}^{m} (-k)_{n_i} \frac{a^k}{k!}. \tag{2.19}$$

Remarque. Lorsque tous les n_i sont égaux à 1, la formule (2.19) se réduit à une formule due à Touchard (cf. [To]); si, de plus, $a = 1$, le polynôme $PC(\mathbf{n}; a)$ devient le nombre de Bell. On retrouve ainsi une formule bien connue due à Dobinski (cf. [Do]).

Démonstration. Il suffit de vérifier que le membre de droite de (2.19) satisfait la même fonction génératrice (2.15) de $PC(\mathbf{n}; a)$. Ceci est évident en appliquant la formule binomiale. □

2.4. Relations entre polynômes générateurs de permutations et de partitions. On se propose ici de trouver des relations simples entre les polynômes $D(\mathbf{n}; \beta, \gamma)$ et $P(\mathbf{n}; \beta, \gamma)$. Le théorème suivant joue un rôle important dans l'interprétation des $\mathcal{M}(\mathbf{n}; \beta, c)$ et

$\mathscr{K}(\mathbf{n}; N, p)$. Pour chaque ensemble $C_{\mathbf{n}}$ associé à la suite $\mathbf{n}=(n_1, \cdots, n_m)$, on pose

$$A_i=\{(i, j): n_1+\cdots+n_{i-1}+1 \leqq j \leqq n_1+\cdots+n_i\}$$

pour $i=1, \cdots, m$, qui est le sous-ensemble des éléments de couleur i de $C_{\mathbf{n}}$. Soit π une permutation de $C_{\mathbf{n}}$, $(\chi(j), j)$ est dit *point fixe* de π si $\chi(j)=\chi(\pi(j))$. On désigne par Fix π l'ensemble des points fixes de π. On établit d'abord le lemme suivant.

LEMME 4. *Soient* $T \subseteq C_{\mathbf{n}}$ *et* $|T \cap A_i|=n_i-k_i$, $i=1, \cdots, m$, *on a alors*

$$\sum_{\mathrm{Fix}\, \pi \supseteq T} w(\pi)=P(\mathbf{k}; \beta, \gamma) \prod_{i=1}^{m}(\beta+k_i)_{n_i-k_i}. \tag{2.20}$$

Remarque. Ce lemme est déjà donné par Foata et Zeilberger lorsque $\gamma=1$ [Fo-Ze]. La démonstration donnée ici est essentiellement la même que celle donnée par ces auteurs.

Démonstration. Posons $T_i=T \cap A_i$ pour $i=1, \cdots, m$. D'après le Lemme 1, le membre de droite de (2.20) est la fonction génératrice du produit $\mathscr{P}(\mathbf{k}) \times \prod_{i=1}^{m} \mathrm{Inj}\,(T_i, A_i)$ par le poids w. Donc pour démontrer le lemme, il suffit d'établir une bijection $\pi \rightarrow(\sigma, \tau_1, \cdots, \pi_m)$, conservant le poids, de $\mathscr{P}(\mathbf{n})$ sur ce produit. Pour cet effet, on part d'une permutation π de $C_{\mathbf{n}}$, fixant tous les éléments de T. On écrit π sous forme de produit de cycles et l'on enlève tous les éléments de T dans chaque cycle. Ce qui reste est alors une permutatoin de $C_{\mathbf{n}} \backslash T$ écrite en produit de cycles. Evidemment, σ a les mêmes excédances de π. On définit ensuite π_i par la restriction de π sur T_i, qui est évidemment une injection de T_i dans A_i. On voit facilement que le nombre totale de cycles de $\sigma, \pi_1, \cdots, \pi_m$ est égal à cyc π. D'autre part, l'application inverse peut se construire immédiatement. □

A l'aide du Lemme 4, on démontre le théorème suivant.

THÉORÈME 3. *On a l'identité*

$$D(\mathbf{n}; \beta, \gamma)=\sum_{k_1 \geqq 0} \cdots \sum_{k_m \geqq 0} P(\mathbf{k}; \beta, \gamma) \prod_{i=1}^{m}(-1)^{n_i-k_i}\binom{n_i}{k_i}(\beta+k_i)_{n_i-k_i}. \tag{2.21}$$

Démonstration. D'après le principe d'inclusion-exclusion, on obtient immédiatement

$$D(\mathbf{n}; \beta, \gamma)=\sum_{T \subseteq C_{\mathbf{n}}}(-1)^{|T|} \sum_{\mathrm{Fix}\, \pi \supseteq T} w(\pi).$$

Par ailleurs, d'après le lemme précédent, pour tout sous-ensemble T de $C_{\mathbf{n}}$ tel que $|T \cap A_i|=n_i-k_i$ $(i=1, \cdots, m)$, on a

$$(-1)^{|T|} \sum_{\mathrm{Fix}\, \pi \supseteq T} w(\pi)=P(\mathbf{k}; \beta, \gamma) \prod_{i=1}^{m}(-1)^{n_i-k_i}(\beta+k_i)_{n_i-k_i}.$$

Il y a évidemment $\prod_{i=1}^{m}\binom{n_i}{k_i}$ de tels sous-ensembles. D'où il résulte (2.21). □

Remarque. Au lieu d'utiliser le principe d'inclusion-exclusion, l'identité (2.21) peut s'établir d'une façon "plus combinatoire," à savoir, construisant une involution sur un ensemble adéquat avec un poids approprié associé à chaque élément. On peut se reporter à l'article de Foata et Zeilberger [Fo-Ze1] pour cette démonstration lorsque $r=1$. Enfin, on peut aussi démontrer (2.21) en identifiant les fonctions génératrices de ses deux membres. Il suffit, en fait, de remplacer $P(\mathbf{k}; \beta, \gamma)$ par la formule (2.12) et d'utiliser la formule (2.9).

On a de même une évaluation analogue pour les polynômes générateurs de partitions. Le théorème suivant établit une relation entre les polynômes $PCS(\mathbf{n}; a)$ et

$PC(\mathbf{n}; a)$. On peut le considérer comme l'analogue du Théorème 3 dans le contexte des partitions colorées.

THÉORÈME 4. *On a l'identiteé*:

$$PCS(\mathbf{n}; a) = \sum_{k_1 \geqq 0} \cdots \sum_{k_m \geqq 0} PC(\mathbf{k}; a) \prod_{i=1}^{m} (-1)^{n_i - k_i} \binom{n_i}{k_i} a^{n_i - k_i}. \tag{2.22}$$

Démonstration. La démonstration de ce théorème est tout à fait semblable à celle du Théorème 3, à ceci près que nous devons remplacer les termes *permutations* par *partitions colorées* et *points fixes* par *points isolés.* Nous nous dispensons donc de la reproduire ici. □

Comme signalé dans la remarque précédente, on peut aussi démontrer (2.22) en identifiant les fonctions génératrices des deux membres.

2.5. Évaluation de polynômes générateurs pour certains cas particuliers. Lorsqu'on se restreint aux suites $\mathbf{n} = (n_1, n_2, n_3)$, les polynômes $PCS(\mathbf{n}; a)$ et $D(\mathbf{n}; \beta, \gamma)$ prennent des formes remarquables données dans les propositions suivantes.

PROPOSITION 1. *On a l'identité*:

$$PCS(n_1, n_2, n_3; a) = \sum_{s \geqq 0} \frac{n_1! n_2! n_3! a^s}{(s-n_1)!(s-n_2)!(x-n_3)!(n_1+n_2+n_3-2s)!}. \tag{2.23}$$

Démonstration. On développe les fonctions symétriques élémentaires e_1, e_2, e_3 en les x_i, puis on identifie les coefficients des monômes $x_1^{n_1} x_2^{n_2} x_3^{n_3}$ dans les deux membres de (2.15). □

PROPOSITION 2. *On a l'identité*:

$$D(n_1, n_2, n_3; \beta, \gamma) = \sum_{s \geqq 0} \frac{n_1! n_2! n_3! (1+\gamma)^{n_1+n_2+n_3-2s}}{(s-n_1)!(s-n_2)!(s-n_3)!(n_1+n_2+n_3-2s)!} \gamma^s (\beta)_s. \tag{2.24}$$

Remarque. On note qu'en faisant $\beta = 1$ dans l'identité précédente, le calcul du membre de gauche devient évident, à savoir

$$n_1! n_2! n_3! \sum_s \binom{n_1}{s-n_2} \binom{n_2}{s-n_3} \binom{n_3}{s-n_1} \gamma^s,$$

et on retrouve la formule classique de Pfaff-Saalschütz (cf. [Ba]) suivante:

$$\sum_{s \geqq 0} \binom{n_1}{s-n_2} \binom{n_2}{s-n_3} \binom{n_3}{s-n_1} \gamma^s = \sum_{s \geqq 0} \frac{s!(1+\gamma)^{n_1+n_2+n_3-2s} \gamma^s}{(s-n_1)!(s-n_2)!(s-n_3)!(n_1+n_2+n_3-2s)!}.$$

C'est exactement sous cette forme que Good [Go] avait donné sa version de la formule de Pfaff-Saalschütz en appliquant le "Master Theorem" de MacMahon et que Foata [Fo1] l'avait démontrée combinatoirement.

Remarque. Les deux Propositions 1 et 2 peuvent être aussi établies de façon authentiquement combinatoire. Cette approche fait l'objet d'un article séparé [Ze1].

Démonstration. Lorsque $m = 3$, la seconde sommation dans le membre de droite de (2.13) se réduit à une seule partition

$$\lambda = (2^{3s-n_1-n_2-n_3} 3^{n_1+n_2+n_3-2s}).$$

D'autre part, le calcul de $a_{\lambda\mu}$ (avec $\mu = (n_1, n_2, n_3)^*$) est facile. On trouve bien

$$a_{\lambda\mu} = \frac{(3s-n_1-n_2-n_3)!}{(s-n_1)!(s-n_2)!(s-n_3)!}.$$

□

COROLLAIRE 5. *Lorsque* $n_3 = 0$, *on a*:

$$D(n_1, n_2; \beta, \gamma) = n_1!\, \gamma^{n_1} (\beta)_{n_1} \delta_{n_1 n_2}, \tag{2.25}$$

$$PCS(n_1, n_2; a) = n_1!\, a^{n_1} \delta_{n_1 n_2}. \tag{2.26}$$

Ces deux identités, qui sont immédiates à partir de (2.23) et de (2.24), serviront à (re)démontrer les propriétés d'orthogonalité des polynômes de Meixner, Krawtchouk, d'une part, et de Charlier, d'autre part.

Notons, enfin, les évaluations de $D(\mathbf{n}; \beta, \gamma)$ et $PCS(\mathbf{n}; a)$ dans le cas m quelconque, mais $n_1 \geqq n_2 + \cdots + n_m$.

PROPOSITION 3. *On a*:

$$D(\mathbf{n}; \beta, \gamma) = \begin{cases} n_1!\, \gamma^{n_1} (\beta)_{n_1} & si\ n_1 = n_2 + \cdots + n_m, \\ 0 & si\ n_1 > n_2 + \cdots + n_m, \end{cases} \tag{2.27}$$

$$PCS(\mathbf{n}; a) = \begin{cases} n_1!\, a^{n_1} & si\ n_1 = n_2 + \cdots + n_m, \\ 0 & si\ n_1 > n_2 + \cdots + n_m. \end{cases} \tag{2.28}$$

Démonstration. Il n'existe pas de dérangements colorés lorsque $n_1 > n_2 + \cdots + n_m$. Lorsqu'on a l'égalité, chaque dérangement coloré π détermine, de façon unique, un ensemble S_π de n_1 *arcs*:

$$[x \to \pi(x)] \qquad (x = (1,1), \cdots, (1, n_1)),$$

ayant chacun une excédance et une permutation π' de S_π donnée par

$$\pi' : [x \to \pi(x)] \mapsto [\pi^2(x) \to \pi(\pi^2(x))] \qquad (x = (1,1), \cdots, (1, n_1)).$$

On remarque, par ailleurs, que π' a le même nombre de cycles que π. Cette application est évidemment bijective. La formule $n_1!(\beta)_{n_1} \gamma^{n_1}$ compte donc d'abord le nombre de façons de construire n_1 arcs, puis le nombre de permutations de ces arcs suivant le nombre de cycles, enfin impose une excédance à chaque arc. La formule (2.28) est évidente. □

Remarque. Lorsque $m = 2$, on retrouve le Corollaire 5.

3. Polynômes orthogonaux.

3.1. Polynômes de Meixner. On est prêt maintenant à interpréter la fonctionnelle $\mathcal{M}(\mathbf{n}; \beta, c)$ définie en (1.4). On rappelle que $D(\mathbf{n}; \beta, \gamma)$ est le polynôme générateur introduit en (2.4).

THÉORÈME 5. *On a*:

$$\mathcal{M}(\mathbf{n}; \beta, c) = D(\mathbf{n}; \beta, c^{-1}). \tag{3.1}$$

Remarque. Dans le cas $\beta = 1$, cette identité a été démontrée par Askey et Ismail [As-Is].

Démonstration. Dans (1.4) développant chaque polynôme de Meixner par la formule (1.1), on obtient

$$\mathcal{M}(\mathbf{n}; \beta, c) = \sum_{k_1 \geqq 0} \cdots \sum_{k_m \geqq 0} F(\mathbf{k}; \beta, c) \prod_{i=1}^{m} (-1)^{n_i - k_i} \binom{n_i}{k_i} (\beta + k_i)_{n_i - k_i}, \tag{3.2}$$

où

$$F(\mathbf{k}; \beta, c) = (1-c)^{\beta} (1 - c^{-1})^{k_1 + \cdots + k_m} \sum_{k \geqq 0} \prod_{i=1}^{m} (-k)_{k_i} \frac{c^k (\beta)_k}{k!}.$$

D'après le Corollaire 1, on a

$$F(\mathbf{k};\beta,c)=P(\mathbf{k};\beta,c^{-1}).$$

Appliquant alors le Théorème 2, on obtient bien (3.1). □

D'après ce théorème, $\mathscr{M}(\mathbf{n};\beta,c)$ est un polynôme de variables β et c^{-1} à coefficients entiers positifs; donc la positivité de $\mathscr{M}(\mathbf{n};\beta,c)$ lorsque $\beta\geqq 0$ et $0<c<1$ devient évidente.

COROLLAIRE 6. *Soient* $\beta\geqq 0$ *et* $0<c<1$, *on a*:

$$\mathscr{M}(\mathbf{n};\beta,c)\geqq 0.$$

En appliquant les propositions données dans la § 2.5, on obtient les corollaires suivants.

COROLLAIRE 7 (Linéarisation). *On a*:

$$(3.3)\quad (n_1,n_2,n_3;\beta,c)=\sum_{s\geqq 0}\frac{n_1!n_2!n_3!(1+c^{-1})^{n_1+n_2+n_3-2s}}{(s-n_1)!(s-n_2)!(s-n_3)!(n_1+n_2+n_3-2s)!}c^{-s}(\beta)_s.$$

Remarque. L'identité (3.3) est déjà donnée par Askey et Gasper [As-Ga] sous une forme équivalente (c'est-à-dire à une constante près).

COROLLAIRE 8 (Orthogonalité). *On a*:

$$(3.4)\qquad \mathscr{M}(n_1,n_2;\beta,c)=n_1c^{-n_1}(\beta)_{n_1}\delta_{n_1n_2}.$$

COROLLAIRE 9. *On a*:

$$(3.5)\qquad \mathscr{M}(\mathbf{n};\beta,c)=\begin{cases}n_1!\,c^{-n_1}(\beta)_{n_1} & \text{si } n_1=n_2+\cdots+n_m,\\ 0 & \text{si } n_1>n_2+\cdots+n_m.\end{cases}$$

3.2. Polynômes de Krawtchouk. L'interprétation de la fonctionnelle $\mathscr{K}(\mathbf{n};N,p)$ définie en (1.5) relève du même modèle que celui utilisé pour $\mathscr{M}(\mathbf{n};\beta,c)$.

THÉORÈME 6. *On a*:

$$(3.6)\qquad \mathscr{K}(\mathbf{n};N,p)=D(\mathbf{n};-N,1-1/p).$$

Démonstration. Dans (1.5) développant chaque polynôme de Krawtchouk par la formule (1.2), on obtient

$$(3.7)\quad \mathscr{K}(\mathbf{n};N,p)=\sum_{k_1\geqq 0}\cdots\sum_{k_m\geqq 0}G(\mathbf{k};N,p)\prod_{i=1}^{m}(-1)^{n_i-k_i}\binom{n_i}{k_i}(-N+k_i)_{n_i-k_i}.$$

où

$$G(\mathbf{k};N,p)=p^{-k_1-\cdots-k_m}\sum_{k\geqq 0}\prod_{i=1}^{m}(-k)_{k_i}\binom{N}{k}p^k(1-p)^{N-k}.$$

A l'aide du Corollaire 1, on vérifie facilement que l'on a:

$$G(\mathbf{k};N,p)=P(\mathbf{k};-N,1-1/p).$$

Appliquant alors le Théorème 3, on obtient bien (3.6). □

Le Théorème 6 affirme que $\mathscr{K}(\mathbf{n};N,p)$ est un polynôme en les variables $(-N)$ et $(1-1/p)$ à coefficients entiers positifs, mais cette fois-ci l'interprétation de $\mathscr{K}(\mathbf{n};N,p)$ ne donne pas d'information sur son signe immédiatement. En fait, il y a deux cas à considérer, comme nous allons l'indiquer après l'énoncé du prochain corollaire. Celui-ci donne une formule nouvelle de la fonctionnelle $\mathscr{K}(\mathbf{n};N,p)$ dans le cas $\mathbf{n}$ quelconque.

COROLLAIRE 10. *On a*:

$$(3.8)\qquad \mathscr{K}(\mathbf{n};N,p)=\prod_{i=1}^{m}n_i!\sum_{s\geqq 0}(-N)_s(p-1)^s\sum_{\lambda}a_{\lambda\mu}\prod_{i=2}^{m}\frac{(1-(1-1/p)^{i-1})^{k_i}}{k_i!},$$

où $\mu = \mathbf{n}^*$ *et* λ *varie dans l'ensemble des partitions* $(1^0 2^{k_2} \cdots m^{k_m})$ *de l'entier* $(n_1 + \cdots + n_m)$ *satisfaisant* $k_2 + \cdots + k_m = s$ *et où* $a_{\lambda\mu}$ *est donné par* $e_\lambda = \sum_\mu a_{\lambda\mu} m_\mu$.

Démonstration. On reporte (3.6) dans (2.14) et on trouve bien (3.8). □

COROLLAIRE 11. *On a les deux inégalités*:

$$\mathcal{K}(\mathbf{n}; N, p) \geqq 0, \qquad \tfrac{1}{2} \leqq p < 1, \tag{3.9}$$

$$(-1)^n \mathcal{K}(\mathbf{n}; N, p) \geqq 0, \qquad 0 < p < \tfrac{1}{2}; \tag{3.10}$$

où n désigne la somme des entiers n_1, n_2, *et* n_3.

Remarque. L'approche de Dunkl et Ramirez [Du-Ra] en termes de caractères de groupes n'a permis à ces auteurs que d'obtenir l'identité (3.8) dans le cas $m = 3$ et pour $\frac{1}{2} \leqq p < 1$. Ils n'ont donc pu établir que (3.9) toujours dans le cas $m = 3$.

Démonstration. Ces inégalités sont une conséquence de l'identité (3.8). D'abord le facteur $(-N)_s (p-1)^s$ est toujours positif pour $0 < p < 1$. Ensuite, lorsque $\frac{1}{2} \leqq p < 1$, on a $-1 \leqq 1 - (1/p) < 0$ et tous les termes du membre de droite de (3.8) sont positifs.

Lorsque $0 < p < \frac{1}{2}$, on a $1 - (1/p) < -1$. Par conséquent, le signe de $\prod_{i=2}^m (1 - (1-1/p)^{i-1})^{k_i}$ est égal à $(-1)^{s(m)}$, où $s(m)$ est la somme des k_i avec *i impair*. Comme $2k_2 + \cdots + mk_m = n_1 + \cdots + n_m$, on a:

$$s(m) \equiv n_1 + \cdots + n_m \pmod 2,$$

qui est indépendant de λ. On en tire immédiatement (3.10). □

Les corollaires suivants sont des conséquences immédiates des propositions correspondantes établies dans la § 2.5.

COROLLAIRE 12 (Linéarisation). *On a*:

$$\mathcal{K}(n_1, n_2, n_3; N, p) = \sum_{s \geqq 0} \frac{n_1! n_2! n_3! (2 - 1/p)^{n_1+n_2+n_3-2s} (1-1/p)^s}{(s-n_1)!(s-n_2)!(s-n_3)!(n_1+n_2+n_3-2s)!} (-N)_s.$$

Remarque. On doit à Eagleson [Ea] d'avoir calculé la fonction génératrice des $\mathcal{K}(n_1, n_2, n_3; N, p)$ et à Askey et Gasper [As-Ga], d'avoir su déduire du calcul d'Eagleson l'identité du Corollaire 12.

COROLLAIRE 13 (Orthogonalité). *On a*:

$$\mathcal{K}(n_1, n_2; N, p) = n_1! (1 - (1/p))^{n_1} (-N)_{n_1} \delta_{n_1 n_2}.$$

COROLLAIRE 14. *On a*:

$$\mathcal{K}(\mathbf{n}; N, p) = \begin{cases} n_1! (1-1/p)^{n_1} (-N)_{n_1} & \text{si } n_1 = n_2 + \cdots + n_m, \\ 0 & \text{si } n_1 > n_2 + \cdots + n_m. \end{cases}$$

3.3. Polynômes de Charlier. L'interprétation de la fonctionnelle $\mathscr{C}(\mathbf{n}; a)$, qui est définie en (1.6), est liée à la notion de *partition colorée*. On rappelle ici que $PCS(\mathbf{n}; a)$ est le polynôme générateur des partitions colorées sans points fixes introduit dans la § 2.3.

THÉORÈME 7. *On a*:

$$\mathscr{C}(\mathbf{n}; a) = PCS(\mathbf{n}; a). \tag{3.11}$$

Démonstration. Dans (1.6) développons chaque polynôme de Charlier par la formule (1.3); on obtient

$$\mathscr{C}(\mathbf{n}; a) = \sum_{k_1 \geqq 0} \cdots \sum_{k_m \geqq 0} T(\mathbf{k}; a) \prod_{i=1}^m (-1)^{n_i - k_i} \binom{n_i}{k_i} a^{n_i - k_i},$$

où

$$T(\mathbf{k};a)=(-1)^{k_1+\cdots+k_m}e^{-a}\sum_{k\geqq 0}\prod_{i=1}^{m}(-k)_{k_i}a^k/k!.$$

Par ailleurs, d'après le Corollaire 4, on a

$$T(\mathbf{k};a)=PC(\mathbf{k};a).$$

Appliquant alors le Théorème 4, on trouve bien (3.11). □

Le Théorème 7 montre que $\mathscr{C}(\mathbf{n};a)$ est un polynôme en la variable a, à coefficients entiers positifs. La positivité de $\mathscr{C}(\mathbf{n};a)$ devient alors évidente lorsque $a>0$.

COROLLAIRE 15. *Soit* $a>0$, *on a*:

$$\mathscr{C}(\mathbf{n};a)\geqq 0.$$

Les corollaires suivants sont des conséquences immédiates des propositions correspondantes établies dans la § 2.5.

COROLLAIRE 16 (Linéarisation). *On a*:

$$\mathscr{C}(n_1,n_2,n_3;a)=\sum_{s\geqq 0}\frac{n_1!n_2!n_3!a^s}{(s-n_1)!(s-n_2)!(s-n_3)!(n_1+n_2+n_3-2s)!}.$$

COROLLAIRE 17 (Orthogonalité). *On a*:

$$\mathscr{C}(n_1,n_2;a)=n_1!a^{n_1}\delta_{n_1n_2}.$$

COROLLAIRE 18. *On a*:

$$\mathscr{C}(\mathbf{n};a)=\begin{cases}n_1!a^{n_1} & si\ n_1=n_2+\cdots+n_m,\\ 0 & si\ n_1>n_2+\cdots+n_m.\end{cases}$$

4. Passage aux limites entre les fonctionnelles. On se rappelle que le tableau d'Askey et Wilson [As-Wi], détaillé par Labelle [La], classe les polynômes orthogonaux classiques d'après leur hiérarchie hypergéométrique. Nous reproduisons dans le tableau 1 la partie située au-dessous des polynômes de Krawtchouk et Meixner.

Une flèche va du polynôme P au polynôme Q, si l'expression analytique de Q peut être obtenue de celle de P par un passage à la limite approprié. Par exemple, nous avons

$$2^n n!\lim_{\beta\to\infty}\beta^{-n}L_n^{(\beta^2/2)}(\beta^2/2-\beta x)=H_n(x), \tag{4.1}$$

où $H_n(x)$ $(n\geqq 0)$ sont les polynômes d'Hermite définis comme suit:

$$\sum_{n\geqq 0}H_n(x)\frac{t^n}{n!}=\exp(2xt-t^2), \tag{4.2}$$

ou

$$H_n(x)=n!\sum_{k\geqq 0}\frac{(-1)^k(2x)^{n-2k}}{(n-2k)!k!}\qquad(n\geqq 0). \tag{4.3}$$

Rappelons que les polynômes de Laguerre $L_n^{(\alpha)}(x)$ sont définis par

$$\sum_{n\geqq 0}L_n^{(\alpha)}(x)u^n=(1-u)^{-\alpha-1}\exp\frac{-xu}{1-u}, \tag{4.4}$$

ou

$$n!L_n^{(\alpha)}(x)=\sum_{k\geqq 0}(-1)^k\binom{n}{k}(\alpha+1+k)_{n-k}x^k\qquad(n\geqq 0). \tag{4.5}$$

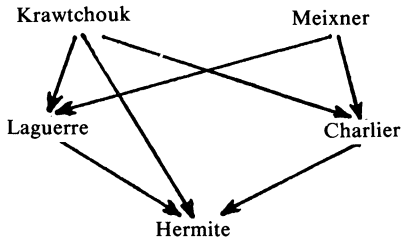


TABLEAU 1

Comme signalé par Foata [Fo3] (voir aussi [La-Ye]), les interprétations combinatoires des polynômes apparaissant dans le diagramme précédent sont connues et compatibles, en ce sens que toutes les formules de passage ont des démonstrations simples dans la géométrie de ces modèles.

Introduisons maintenant les fonctionnelles des polynômes de Laguerre et d'Hermite:

$$\mathscr{L}(\mathbf{n};\alpha)=\frac{1}{\Gamma(\alpha+1)}\left(\prod_{j=1}^{m}(-1)^{n_i}n_i!\right)\int_0^\infty\left(\prod_{i=1}^{m}L_{n_i}^{(\alpha)}(x)\right)x^\alpha e^{-x}\,dx, \tag{4.6}$$

$$\mathscr{H}(\mathbf{n})=(2^{n_1+\cdots+n_m}\pi)^{-1/2}\int_{-\infty}^{\infty}\left(\prod_{i=1}^{m}H_{n_i}(x)\right)e^{-x^2}\,dx. \tag{4.7}$$

Les interprétations combinatoires de ces deux fonctionnelles sont déjà données respectivement par Foata et Zeilberger [Fo-Ze1] et Azor, Gillis, et Victor [Az-Gi-Vi]. On se propose ici de redémontrer d'abord l'interprétation combinatoire de $\mathscr{L}(\mathbf{n};\alpha)$ à partir de celle de $\mathscr{M}(\mathbf{n};\beta,\gamma)$ et l'interprétation de $\mathscr{H}(\mathbf{n})$ à partir de celle de $\mathscr{L}(\mathbf{n};\alpha)$. On démontre aussi que les interprétations combinatoires des fonctionnelles des polynômes apparaissant dans le diagramme sont compatibles dans le même sens que ci-dessus.

THÉORÈME 8 (Foata et Zeilberger). *On a*:

$$\mathscr{L}(\mathbf{n};\alpha)=D(\mathbf{n};\alpha+1,1). \tag{4.8}$$

Démonstration. Dans (4.6) développant chaque polynôme de Laguerre $L_{n_i}^{(\alpha)}(x)$ par (4.5) et intégrant terme par terme en utilisant le fait que $(1/\Gamma(\alpha+1)\int_0^\infty e^{-x}x^{n+\alpha}\,dx=(\alpha+1)_n$ (cf. [Fo-Ze]), on obtient

$$\mathscr{L}(\mathbf{n};\alpha)=\sum_{k_1\geqq 0}\cdots\sum_{k_m\geqq 0}(\alpha+1)_{k_1+\cdots+k_m}\prod_{i=1}^{m}(-1)^{n_i-k_i}\binom{n_i}{k_i}(\alpha+1+k_i)_{n_i-k_i}. \tag{4.9}$$

En comparant (4.9) avec (3.2) et en appliquant le Corollaire 2, on a immédiatement:

$$\mathscr{L}(\mathbf{n};\alpha)=\lim_{c\to 1}\mathscr{M}(\mathbf{n};\alpha+1,c). \tag{4.10}$$

D'autre part, l'identité $\mathscr{D}(\mathbf{n};\alpha+1,1)=\lim_{c\to 1}\mathscr{D}(\mathbf{n};\alpha+1,c^{-1})$ est évidente. En appliquant le Théorème 5, on a (4.8). □

On appelle *involution dérangée* toute involution de $C_{\mathbf{n}}$, qui est en même temps un dérangement de $C_{\mathbf{n}}$. On désigne par Invd $(\mathbf{n})$ l'ensemble des involutions dérangées de $C_{\mathbf{n}}$.

THÉORÈME 9 (Azor, Gillis, et Victor). *On a*:

$$\mathscr{H}(\mathbf{n})=|\mathrm{Invd}\,(\mathbf{n})|. \tag{4.11}$$

Démonstration. C'est un exercice élémentaire d'analyse de vérifier

$$\lim_{\beta\to\infty}(\sqrt{2}\,\beta^{-1})^{n_1+\cdots+n_m}\mathscr{L}(\mathbf{n};\beta^2/2)=(-1)^{n_1+\cdots+n_m}\mathscr{H}(\mathbf{n}), \tag{4.12}$$

à l'aide de (4.2). Par ailleurs, on a

$$(\sqrt{2}\,\beta^{-1})^{n_1+\cdots+n_m}D(\mathbf{n};\beta^2/2+1,1)=\sum_{\pi\in\mathscr{D}(\mathbf{n})}(\beta^2/2+1)^{\mathrm{cyc}\,\pi}(\sqrt{2}\,\beta^{-1})^{n_1+\cdots+n_m},$$

lorsque $\beta\to\infty$, tous les termes de la somme tendent vers zéro, sauf ceux qui satisfont $2\,\mathrm{cyc}\,\pi=\sum_{i=1}^{m}n_i$, c'est-à-dire ceux qui correspondent aux involutions dérangées de $C_{\mathbf{n}}$; or ces termes tendent vers 1. Par conséquent, d'après le Théorème 8, on a

$$(-1)^{n_1+\cdots+n_m}\mathscr{H}(\mathbf{n})=|\mathrm{Invd}\,(\mathbf{n})|.$$

Comme il n'existe pas d'involution dérangée lorsque $n_1+\cdots+n_m$ est impaire, on a donc

$$\mathscr{H}(\mathbf{n})=|\mathrm{Invd}\,(\mathbf{n})|. \qquad \square$$

Les interprétations combinatoires (Théorème 5-9) des fonctionnelles définies dans cet article permettent de trouver facilement les passages au limite entre ces fonctionnelles.

THÉORÈME 10. *On a*:

$$\lim_{c\to 1}\mathscr{M}(\mathbf{n};\alpha+1,c)=\mathscr{L}(\mathbf{n};\alpha), \tag{4.13}$$

$$\lim_{\beta\to\infty}\left(\frac{a}{\beta}\right)^{n_1+\cdots+n_m}\mathscr{M}\left(\mathbf{n};\beta,\frac{a}{\beta}\right)=\mathscr{C}(\mathbf{n};a), \tag{4.14}$$

$$\lim_{\beta\to\infty}\left(\frac{\sqrt{2}}{\beta}\right)^{n_1+\cdots+n_m}\mathscr{L}\left(\mathbf{n};\frac{\beta^2}{2}\right)=\mathscr{H}(\mathbf{n}), \tag{4.15}$$

$$\lim_{a\to\infty}\left(\frac{1}{\sqrt{a}}\right)^{n_1+\cdots+n_m}\mathscr{C}(\mathbf{n};a)=\mathscr{H}(\mathbf{n}), \tag{4.16}$$

$$\lim_{N\to\infty}\left(-\frac{a}{n}\right)^{n_1+\cdots+n_m}\mathscr{K}\left(\mathbf{n};N,\frac{a}{N}\right)=\mathscr{C}(\mathbf{n};a), \tag{4.17}$$

$$\lim_{N\to\infty}\left(\frac{1}{\sqrt{N}}\right)^{n_1+\cdots+n_m}\mathscr{K}\left(\mathbf{n};N,\frac{1}{2}\right)=\mathscr{H}(\mathbf{n}). \tag{4.18}$$

Démonstration. L'identité (4.13) (respectivement, (4.15)) est, en fait, déjà démontrée dans la démonstration du Théorème 8 (respectivement, 9), si l'on part d'abord de l'interprétation de (4.8) (respectivement, (4.11)). Les autres identités sont démontrées de la façon analogue, à l'aide de leurs interprétations. $\square$

5. Appendice: quelques remarques sur le Corollaire 2. Dans la § 2.2, on a établi, de façon relativement élaborée, le résultat analytique suivant (Corollaire 2):

$$\lim_{\gamma\to 1}(1-\gamma)^{\beta}(1-\gamma)^{n_1+\cdots+n_m}\sum_{k\geqq 0}\prod_{i=1}^{m}(-k)_{n_i}\frac{\gamma^{-k}(\beta)_k}{k!}=(\beta)_{n_1+\cdots+n_m}, \tag{5.1}$$

où $n_1,\cdots,n_m$ sont m entiers positifs. Il est intéressant de transcrire ce résultat en termes de fonctions hypergéométriques et de voir si les identités classiques sur ces fonctions permettent de retrouver (5.1). Rappelons que si r et s sont deux entiers positifs, $(a_1,\cdots,a_r)$ et $(b_1,\cdots,b_s)$ sont deux suites de paramètres (réels ou complexes), la fonction hypergéométrique ${}_rF_s$ est donnée par (cf. [Ba]):

$${}_rF_s\left(\begin{matrix}a_1,\cdots,a_r\\ b_1,\cdots,b_s\end{matrix};x\right)=\sum_{n\geqq 0}\frac{(a_1)_n\cdots(a_r)_n}{(b_1)_n\cdots(b_s)_n}\frac{x^n}{n!}.$$

Prenons $n_1 = \max\{n_1, n_2, \cdots, n_m\}$ et récrivons le membre de droite de (5.1) à l'aide de ces fonctions. On obtient l'énoncé équivalent:

$$\lim_{\gamma\to 1} \gamma^{n_2+\cdots+n_m}(1-\gamma^{-1})^{\beta+n_1+\cdots+n_m} \frac{(n_1!)^{m-1}(\beta)_{n_1}}{\prod_{i=1}^{m}(n_1-n_i)!} \times {}_mF_{m-1}\left(\begin{matrix}\beta+n_1, n_1+1, \cdots, n_1+1 \\ n_1-n_2+1, n_1-n_3+1, \cdots, n_1-n_m+1\end{matrix}; \gamma^{-1}\right) = (\beta)_{n_1+\cdots+n_m}.$$

Si l'on supprime le facteur $\lim_{\gamma\to 1} \gamma^{n_2+\cdots+n_m} = 1$ et si l'on pose $x = \gamma^{-1}$, on est conduit à l'énoncé

$$\lim_{x\to 1} (1-x)^{\beta+n_1+\cdots+n_m} \frac{(n_1!)^{m-1}(\beta)_{n_1}}{\prod_{i=1}^{m}(n_1-n_i)!} \times {}_mF_{m-1}\left(\begin{matrix}\beta+n_1, n_1+1, \cdots, n_1+1 \\ n_1-n_2+1, n_1-n_3+1, \cdots, n_1-n_m+1\end{matrix}; x\right) = (\beta)_{n_1+\cdots+n_m},$$

ou, de façon équivalente, à:

$$\lim_{x\to 1} (1-x)^{\beta+n_1+\cdots+n_m} \times {}_mF_{m-1}\left(\begin{matrix}\beta+n_1, n_1+1, \cdots, n_1+1 \\ n_1-n_2+1, n_1-n_3+1, \cdots, n_1-n_m+1\end{matrix}; x\right) = \frac{\prod_{i=1}^{m}(n_1-n_i)!}{(n_1!)^{m-1}(\beta)_{n_1}} (\beta)_{n_1+\cdots+n_m} \left(= \frac{(\beta+n_1)_{n_2+\cdots+n_m}}{\prod_{i=2}^{m}(n_1-n_i+1)_{n_i}}\right). \tag{5.2}$$

Pour $m \leqq 2$, l'identité (5.2) peut être vérifiée directement à l'aide de la formule binomiale, d'Euler et de Gauss. Lorsque $m = 1$, l'énoncé (5.2), avant le passage à la limite, se réduit à

$$(1-x)^{\beta+n_1}{}_1F_0\left(\begin{matrix}\beta+n_1 \\ -\end{matrix}; x\right) = 1,$$

qui n'est autre que l'identité binomiale.

Pour $m = 2$, on peut appliquer la transformation d'Euler (cf. [Ba, p. 8]) à la fonction ${}_mF_{m-1} = {}_2F_1$,

$${}_2F_1\left(\begin{matrix}\beta+n_1, n_1+1 \\ n_1-n_2+1\end{matrix}; x\right) = (1-x)^{-\beta-n_1-n_2}{}_2F_1\left(\begin{matrix}-n_2-\beta+1, -n_2 \\ n_1-n_2+1\end{matrix}; x\right).$$

Le premier membre de (5.2) devient donc:

$${}_2F_1\left(\begin{matrix}-n_2-\beta+1, -n_2 \\ n_1-n_2+1\end{matrix}; x\right).$$

On peut alors appliquer la formule de Gauss (cf. [Ba, p. 6]) lorsque x tend vers 1.

$$\lim_{x\to 1} {}_2F_1\left(\begin{matrix}-n_2-\beta+1, -n_2 \\ n_1-n_2+1\end{matrix}; x\right) = \frac{\Gamma(n_1-n_2+1)\Gamma(n_1+n_2+\beta)}{\Gamma(n_1+1)\Gamma(n_1+\beta)} = \frac{(n_1-n_2)!(\beta)_{n_1+n_2}}{n_1!(\beta)_{n_1}}.$$

Ce qui établit (5.2) dans le cas $m = 2$.

Comme les transformations utilisées ci-dessus (binomiale, Euler, Gauss) n'ont pas de généralisations analogues dans le cas m quelconque, il semble qu'une transformation appropriée reste à trouver. En fait, on a seulement besoin de connaître une formule asymptotique pour la série ${}_mF_{m-1}$ ci-dessus lorsque x tend vers 1.

Remerciement. L'auteur tient à remercier M. D. Foata pour son aide et ses suggestions durant toute la préparation de ce travail.

BIBLIOGRAPHIE

[As] R. ASKEY, *Linearization of the product of orthogonal polynomials*, in Problems in Analysis (A symposium in honor of Salomon Bochner), Robert C. Gunning, ed., Princeton University Press, Princeton, NJ, 1970, pp. 131-138.

[As2] ———, *Orthogonal and Special Functions*, CBMS-NSF Regional Conference Series in Applied Mathematics, 21, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1975.

[As-Ga] R. ASKEY AND G. GASPER, *Convolution structures for Laguerre polynomials*, J. d'Analyse Math., 31 (1977), pp. 46-48.

[As-Is] R. ASKEY AND M. E. H. ISMAIL, *Permutation problems and special functions*, Canad. J. Math., 28 (1976), pp. 853-874.

[As-Wi] R. ASKEY AND J. A. WILSON, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 318, Providence, RI, 1985.

[Az-Gi-Vi] R. AZOR, J. GILLIS, AND J. D. VICTOR, *Combinatorial application of Hermite polynomials*, SIAM J. Math. Anal., 13 (1982), pp. 879-890.

[Ba] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, Cambridge, 1935.

[Ca-Fo] P. CARTIER AND D. FOATA, *Problèmes combinatoires de permutations et réarrangements*, Lecture Notes in Math. 85, Springer-Verlag, Berlin, 1969.

[Ch] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[Do] G. DOBINSKI, *Summierung der Reihe* $\sum n^m/n!$ *für* $m = 1, 2, 3, 4, 5, \cdots$, Arch. Math. Physik, 61 (1877), pp. 333-336.

[Du-Ra] C. F. DUNKL AND D. E. RAMIREZ, *Krawtchouk polynomials and the symmetrization of hypergroups*, SIAM J. Math. Anal., 5 (1974), pp. 351-366.

[Ea] G. K. EAGLESON, *A characterization theorem for positive definite sequences on the Krawtchouk polynomials*, Austral. J. Statist., 11 (1969), pp. 29-38.

[Ev-Gi] S. EVEN AND J. GILLIS, *Derangements and Laguerre polynomials*, Proc. Cambridge Philos. Soc., 79 (1976), pp. 135-143.

[Fo1] D. FOATA, *Étude algébrique de certains problèmes d'analyse combinatoire et du calcul des probabilités*, Publ. Inst. Statist. Univ. Paris, 14 (1965), pp. 81-241.

[Fo2] ———, *La série génératrice exponentielle dans les problèmes d'énumération*, Presses de l'Université de Montréal, Montréal, 1974.

[Fo3] ———, *Combinatoire des identités sur les polynômes orthogonaux*, Proc. Intern. Congress of Mathematics, Warsaw, August 16-24, 1983, Varsovie, 1983, pp. 1541-1553.

[Fo-La] D. FOATA AND J. LABELLE, *Modèles combinatoires pour les polynômes de Meixner*, Europ. J. Combin., 4 (1983), pp. 305-311.

[Fo-Sch] D. FOATA AND M.-P. SCHÜTZENBERGER, *Théorie géométrique des polynômes eulériens*, Lecture Notes in Math. 138, Springer-Verlag, Berlin, 1970.

[Fo-St] D. FOATA AND V. STREHL, *Combinatorics of Laguerre polynomials, Enumeration and Design*, Waterloo, June-July 1982, D. M. Jackson and S. A. Vanstone, eds., Academic Press, Toronto, 1984, pp. 123-140.

[Fo-Ze1] D. FOATA AND D. ZEILBERGER, *Weighted derangements and Laguerre polynomials*, Actes 8^e Séminaire Lotharingien, Publ. I.R.M.A. Strasbourg, 229/S-08, 1984, pp. 17-25.

[Fo-Ze2] ———, *Laguerre polynomials, weighted derangements and positivity*, SIAM J. Discrete Math., 1 (1988), pp. 425-433.

[Ge] I. GESSEL, *Generalized Rook polynomials and orthogonal polynomials*, exposé oral, Strasbourg, 1987.

[Gi-Je-Ze] J. GILLIS, J. JEDWAB, AND D. ZEILBERGER, *A combinatorial interpretation of the integral of the product of Legendre polynomials*, SIAM J. Math. Anal., 19 (1988), pp. 1455-1461.

[Go] I. J. GOODS, *Proofs of some "binomial" identities by means of MacMahon's "Master Theorem,"* Proc. Cambridge Philos. Soc., 58 (1962), pp. 161–162.

[Ja] D. M. JACKSON, *Laguerre polynomials and derangements*, Proc. Cambridge Philos. Soc., 80 (1976), pp. 213–214.

[La] J. LABELLE, *Tableau d'Askey*, in Polynômes orthogonaux et applications, Bar-le-Duc, 1984, C. Brezinski et al., eds., Lecture Notes in Math. 1171, Springer-Verlag, Berlin, 1985, pp. xxxvi–xxxvii.

[La-Ye] J. LABELLE AND Y. N. YEH, *Combinatorial proofs of some limit formulas involving orthogonal polynomials*, Discrete Math., 79 (1989/90), pp. 77–93.

[Macd] I. MACDONALD, *Symmetric Functions and Hall Polynomials*, Clarendon Press, Oxford, 1979.

[Mac] P. A. MACMAHON, *Combinatory Analysis*, Vol. 1, Cambridge University Press, Cambridge, 1915. (Reprinted by Chelsea, New York, 1955.)

[Mu] T. MUIR, *A treatise on the theory of determinants*, Longmans, London, 1933. (Reprinted by Dover, New York, 1960.)

[Ra] M. RAHMAN, *A non-negative representation of the linearization coefficients of the product of Jacobi polynomials*, Canad. J. Math., 33 (1981), pp. 915–928.

[Ri] J. RIORDAN, *An Introduction to Combinatorial Analysis*, John Wiley, New York, 1959.

[Sa-Vi] M. DE SAINTE-CATHERINE AND G. VIENNOT, *Combinatorial interpretation of integrals of products of Hermite, Laguerre and Tchebycheff polynomials*, in Polynômes orthogonaux et applications, Bar-le-Duc, 1984, C. Brezinski et al., eds., Lectures Notes in Math. 1171, Springer-Verlag, Berlin, 1985, pp. 120–128.

[Sz] G. SZEGÖ, *Orthogonal polynomials*, American Mathematical Society, Providence, RI, 1939 (Amer. Math. Soc. Colloq. Publ., 23), (second printing of the fourth edition, 1978).

[To] J. TOUCHARD, *Nombres exponentiels et nombres de Bernoulli*, Canad. J. Math., 8 (1956), pp. 305–320.

[Vi] G. VIENNOT, *Une théorie combinatoire des polynômes orthogonaux généraux*, Conference Notes, Université du Québec à Montréal, 1984. (In French).

[Ze1] J. ZENG, *Calcul saalschützien des partitions et dérangements colorés*, SIAM J. Discrete Math. 3 (1990), pp. 149–156.

[Ze2] ———, *La β-extension de la formule d'inversion de Lagrange à plusieurs variables*, Stud. Appl. Math., to appear.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1369–1385, November 1990
©1990 Society for Industrial and Applied Mathematics

LOCAL EXISTENCE OF SOLUTIONS OF THE DIRICHLET INITIAL-BOUNDARY VALUE PROBLEM FOR INCOMPRESSIBLE HYPOELASTIC MATERIALS*

MICHAEL RENARDY†

Abstract. Energy methods are used to show the well-posedness of the Dirichlet initial-boundary value problem for incompressible hypoelastic materials and for a related class of viscoelastic fluids.

Key words. hypoelasticity, polymer rheology, local existence

AMS(MOS) subject classifications. 35L20, 35L70, 73C50, 76A10

1. Introduction. The equation of motion for an incompressible continuous medium is

$$\rho\frac{dv}{dt} = \rho\Big(\frac{\partial v}{\partial t} + (v\cdot\nabla)v\Big) = \operatorname{div}\mathbf{T} - \nabla p + f, \tag{1}$$

with the incompressibility constraint

$$\operatorname{div} v = 0. \tag{2}$$

Here v is the velocity, p the pressure, $\mathbf{T}$ the extra Cauchy stress, ρ the density, and f a given body force. Throughout this paper, d/dt denotes the material time derivative, while $\partial/\partial t$ or a dot denotes the Eulerian time derivative.

In addition to equations (1) and (2) we need a constitutive law, which relates $\mathbf{T}$ to the motion. For isotropic elastic materials, the stress $\mathbf{T}$ is a function of the (Finger) strain. By differentiating such a relationship with respect to time, we obtain a linear relationship between the time derivative of the stress and the velocity gradient, with coefficients depending on the strain or, alternatively (if the relationship between stress and strain is invertible), on the stress. Hypoelastic materials [16] were introduced as a natural generalization. In a hypoelastic materials, the material time derivative of the stress is given as a linear function of the velocity gradient, with coefficients depending on the stress. If special integrability conditions are satisfied, then integration with respect to time yields an elastic constitutive law, but in general this is not the case. Hypoelasticity has sometimes been used to model the behavior of solids, but does not seem to have gained much popularity. There is, however, a closely related class of models which is obtained by including a lower-order term in the equation. These models have the form

$$\frac{d}{dt}T_{ij} = A_{ijkl}(\mathbf{T})\frac{\partial v_k}{\partial x_l} + g_{ij}(\mathbf{T}), \tag{3}$$

*Received by the editors May 1, 1989; accepted for publication November 20, 1989. This research was completed while the author was visiting the Centre for Mathematical Analysis, Australian National University. Financial support was provided by the Centre and by National Science Foundation grant DMS-8796241.

†Department of Mathematics and Interdisciplinary Center for Applied Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0123.

where we have adopted the summation convention. Models of the form (3) are widely used in polymer rheology. Although they are not a realistic description of real polymers, they have some qualitative merits, and, due to their relative simplicity, they are popular for numerical simulations. We refer to [8] and [11] for some specific examples of such models; there are many others in the literature. For the purpose of illustration, we cite the Johnson–Segalman model:

$$\frac{d\mathbf{T}}{dt} - \frac{1+a}{2}(\nabla v\mathbf{T} + \mathbf{T}(\nabla v)^T) + \frac{1-a}{2}(\mathbf{T}\nabla v + (\nabla v)^T\mathbf{T}) + \lambda\mathbf{T}$$

$$= \mu(\nabla v + (\nabla v)^T).$$

Here λ and μ are positive constants, and a is a constant which lies between 1 and -1.

We seek solutions of (1)–(3) for $t > 0$ and $x \in \Omega$, where Ω is a bounded domain in $\mathbb{R}^3$. We impose the boundary condition

$$v(x,t) = 0, \qquad x \in \partial\Omega, \tag{4}$$

and initial conditions

$$v(x,0) = v_0(x), \qquad \mathbf{T}(x,0) = \mathbf{T}_0(x). \tag{5}$$

The principles of continuum mechanics impose certain restrictions on the coefficients A_{ijkl} (see [16]). Frame-indifference requires that

$$A_{ijkl}(\mathbf{T}) = \tfrac{1}{2}[\delta_{ik}T_{lj} - \delta_{il}T_{kj} - T_{ik}\delta_{lj} + T_{il}\delta_{kj}] + B_{ijkl}(\mathbf{T}), \tag{6}$$

where B_{ijkl} is symmetric in k and l (and of course also in i and j). Moreover, with $\mathbf{D}$ denoting the symmetric part of the velocity gradient, the function $(\mathbf{D},\mathbf{T}) \mapsto \mathbf{B}(\mathbf{T})\mathbf{D} + \mathbf{g}(\mathbf{T})$ must be isotropic (we shall, however, not use this fact). We shall assume that

$$B_{ijkl} = B_{klij}. \tag{7}$$

The analogous symmetry condition in elasticity follows from the existence of an elastic energy; in the context of hypoelasticity, however, the physical meaning of (7) is not obvious, in particular, (7) does not follow from thermodynamics. We set

$$C_{ijkl} = B_{ijkl} + \tfrac{1}{2}[\delta_{ik}T_{lj} - \delta_{il}T_{kj} - T_{ik}\delta_{lj} - T_{il}\delta_{kj}], \tag{8}$$

so that

$$A_{ijkl} = T_{il}\delta_{kj} + C_{ijkl}, \tag{9}$$

and we observe that C_{ijkl} also satisfies the symmetry condition

$$C_{ijkl} = C_{klij}. \tag{10}$$

As in elasticity, a strong ellipticity condition is essential for the well-posedness of the initial value problem. We require that

$$C_{ijkl}(\mathbf{T})\zeta_i\zeta_k\eta_j\eta_l \geq \kappa(\mathbf{T})|\zeta|^2|\eta|^2 \quad \forall \zeta,\eta \in \mathbb{R}^3, \tag{11}$$

where $\kappa(\mathbf{T}) > 0$. It would suffice to require (11) only for ζ, η with $\zeta \cdot \eta = 0$; however, we can then make it true for all ζ, η by adding a multiple of $\delta_{ij}\delta_{kl}$ to C_{ijkl}; for divergence-free v, this causes no change in (3).

Well-posedness for elastodynamics has been shown only recently. Hughes, Kato and Marsden [7] consider the initial value problem on all of space for a compressible elastic medium. Subsequently, the Dirichlet initial-boundary value problem has been treated by Kato [9], Chen and von Wahl [2], and Dafermos and Hrusa [3]. For incompressible elastic materials, the initial value problem has been solved by Schochet [13] and Ebin and Saxton [4]; the Dirichlet initial-boundary value problem has been treated by Hrusa and Renardy [6]. The proofs in the present paper will to some extent rely on ideas used in [6].

We make the following assumptions of smoothness, ellipticity, and compatibility:

(S1) The domain $\Omega \subset \mathbb{R}^3$ is bounded and $\partial\Omega$ is of class C^5.

(S2) The functions A_{ijkl} and g_{ij} are of class C^4.

(S3) $v_0 \in H^4(\Omega)$, $\mathbf{T}_0 \in H^4(\Omega)$.

(S4) For some $T > 0$, we have $f \in \bigcap_{k=0}^{4} W^{k,1}([0,T]; H^{4-k}(\Omega))$.

(E) The symmetry condition (10) holds and there is a continuous function κ, defined in a neighborhood of the range of $\mathbf{T}_0$ and taking values in $\mathbb{R}^+$, such that (11) holds.

(C1) $\operatorname{div} v_0 = 0$ and $v_0 = 0$ on $\partial\Omega$.

(C2) The initial values of $\dot{v}$, $\ddot{v}$ and $\partial^3 v/\partial t^3$, as determined from the equations, vanish on $\partial\Omega$.

Here the initial value of $\dot{v}$ is obtained by applying the Hodge projection (i.e., the orthogonal projection in $L^2(\Omega)$ onto the subspace of divergence-free vector fields with vanishing normal component on $\partial\Omega$, see e.g., [15]) to equation (1). This eliminates p and everything else is given at $t = 0$. Similarly, we find the initial values of $\ddot{v}$ and $\partial^3 v/\partial t^3$ by differentiating (1) with respect to time and applying the Hodge projection to the differentiated equations.

THEOREM. *Assume that* (S1)–(S4), (E), *and* (C1)–(C2) *hold. Then there is a* $T' \in (0,T]$ *such that problem* (1)–(5) *has a unique solution with the regularity*

$$v \in \bigcap_{k=0}^{4} C^k([0,T']; H^{4-k}(\Omega)); \qquad \mathbf{T} \in \bigcap_{k=0}^{3} C^k([0,T']; H^{3-k}(\Omega)).$$

Remarks.

1. It may appear strange that we require $\mathbf{T}_0 \in H^4(\Omega)$, while the solution $\mathbf{T}$ only lies in $H^3(\Omega)$. However, the assumption $\mathbf{T}_0 \in H^4(\Omega)$ is merely a convenient way of guaranteeing that the initial values of $\partial^k v/\partial t^k$, $k = 1,2,3,4$, lie in $H^{4-k}(\Omega)$. This latter feature is reproduced by the solution.

2. A similar result with a much easier proof can be established for compressible hypoelastic materials. In that case, the initial value of the density must also be prescribed.

3. If more regularity of the data and additional compatibility conditions are assumed, then more regularity of the solution is obtained.

4. It is possible to replace the condition $v = 0$ on $\partial\Omega$ by an inhomogeneous Dirichlet condition, as long as the prescribed velocity is tangential to $\partial\Omega$. If it is not, then the imposition of boundary conditions leading to a well-posed problem becomes a very tricky issue. For a discussion of inflow boundary conditions in the context of steady flows of viscoelastic fluids see [12].

2. Construction of solution. In this section, we define the iteration scheme which is used to construct a solution. First, we apply the operator $d/dt + (\nabla v)^T$ to the equation of motion (1) (the matrix ∇v is defined such that the first index refers to the components of v and superscript T denotes the transposed matrix), and we substitute the right-hand side of (3) for $d\mathbf{T}/dt$. This yields the following equation:

$$\rho\Big[\frac{\partial}{\partial t} + (v\cdot\nabla)\Big]^2 v_i = -\frac{\partial q}{\partial x_i} + A_{ijkl}(\mathbf{T})\frac{\partial^2 v_k}{\partial x_j \partial x_l} + h_i(v, \nabla v, \dot{v}, \mathbf{T}, \nabla\mathbf{T}, f, \nabla f, \dot{f}). \tag{12}$$

Here we have set $q = (\partial/\partial t + (v\cdot\nabla))p$, and

$$\begin{aligned} h_i = &-\rho\frac{\partial v_j}{\partial x_i}\left[\frac{\partial}{\partial t} + (v\cdot\nabla)\right]v_j + \frac{\partial v_j}{\partial x_i}\frac{\partial T_{jk}}{\partial x_k} - \frac{\partial v_k}{\partial x_j}\frac{\partial T_{ij}}{\partial x_k} \\ &+ \frac{\partial}{\partial x_j}g_{ij}(\mathbf{T}) + \Big(\frac{\partial}{\partial x_j}A_{ijkl}(\mathbf{T})\Big)\frac{\partial v_k}{\partial x_l} + \left[\frac{\partial}{\partial t} + (v\cdot\nabla)\right]f_i + \frac{\partial v_j}{\partial x_i}f_j. \end{aligned} \tag{13}$$

We note that because of (2) and (9) we have

$$A_{ijkl}\frac{\partial^2 v_k}{\partial x_j \partial x_l} = C_{ijkl}\frac{\partial^2 v_k}{\partial x_j \partial x_l}. \tag{14}$$

We construct our solution by the following iteration scheme. For given v^n, we determine $\mathbf{T}^n$ from the initial value problem

$$\begin{aligned} &\left[\frac{\partial}{\partial t} + (v^n\cdot\nabla)\right]T^n_{ij} = A_{ijkl}(\mathbf{T}^n)\frac{\partial v^n_k}{\partial x_l} + g_{ij}(\mathbf{T}^n), \\ &\mathbf{T}^n(x,0) = \mathbf{T}_0(x). \end{aligned} \tag{15}$$

Then v^{n+1} is determined from

$$\begin{aligned} &\rho\Big[\frac{\partial}{\partial t} + (v^n\cdot\nabla)\Big]^2 v_i^{n+1} = -\frac{\partial q^{n+1}}{\partial x_i} + C_{ijkl}(\mathbf{T}^n)\frac{\partial^2 v_k^{n+1}}{\partial x_j \partial x_l} \\ &\qquad + h_i(v^n, \nabla v^n, \dot{v}^n, \mathbf{T}^n, \nabla\mathbf{T}^n, f, \nabla f, \dot{f}), \\ &\operatorname{div} v^{n+1} = 0, \quad v^{n+1}|_{\partial\Omega} = 0, \quad v^{n+1}(x,0) = v_0(x), \quad \dot{v}^{n+1}(x,0) = v_1(x). \end{aligned} \tag{16}$$

Here v_1 is the initial value of $\dot{v}$, defined as explained above. While the solution of (15) is relatively easy, the construction of solutions to (16) is quite complicated; it will involve an "inner" iteration an a Galerkin approximation. A rough outline of the procedure is sketched at the end of this section, and details are given in §§3 and 4. Our eventual task is to show that the mapping $\Sigma : v^n \mapsto v^{n+1}$ is a contraction in an appropriate complete metric space. The fixed point of the contraction is the solution we seek.

Let $Z(M, T')$ be the set of all functions $v : \Omega \times [0, T'] \to \mathbb{R}^3$ with the following properties:

$$v \in \bigcap_{k=0}^{4} W^{k,\infty}([0,T']; H^{4-k}(\Omega)), \tag{17}_1$$

$$(17)_2 \qquad \|v\|_{0,4} + \|v\|_{1,3} + \|v\|_{2,2} + \|v\|_{3,1} + \|v\|_{4,0} \leq M,$$

$$(17)_3 \qquad \operatorname{div} v = 0,$$

$$(17)_4 \qquad v|_{\partial\Omega} = 0,$$

$$(17)_5 \qquad v(x,0) = v_0(x), \quad \dot{v}(x,0) = v_1(x), \quad \ddot{v}(x,0) = v_2(x), \quad \frac{\partial^3 v}{\partial t^3}(x,0) = v_3(x).$$

Here $\|\cdot\|_{k,l}$ denotes the norm in $W^{k,\infty}([0,T'];H^l(\Omega))$; later we shall also use the notation $\|\cdot\|_{k,l,p}$ for the norm in $W^{k,p}([0,T'];H^l(\Omega))$. The functions v_1, v_2, and v_3 are the (known) initial values of $\dot{v}$, $\ddot{v}$ and $\partial^3 v/\partial t^3$; they lie in $H^3(\Omega)$, $H^2(\Omega)$, and $H^1(\Omega)$, respectively. On $Z(M,T')$, we define the metric

$$(18) \qquad d(v,\hat{v}) = \|v-\hat{v}\|_{0,3} + \|v-\hat{v}\|_{1,2} + \|v-\hat{v}\|_{2,1} + \|v-\hat{v}\|_{3,0}.$$

LEMMA 1. *If M is chosen large enough, $Z(M,T')$ is not empty. Moreover, the metric d is complete on $Z(M,T')$.*

Proof. It is easy to see that $Z(M,T')$ is complete, and we shall only give the proof that it is not empty. For this, we have to construct a function v which satisfies $(17)_1$ and $(17)_3$–$(17)_5$. Let $a \in H^4(\Omega)$ be a vector field such that on $\partial\Omega$ we have

$$(19) \qquad a = 0, \quad \frac{\partial a}{\partial n} = 0, \quad \frac{\partial^2 a}{\partial n^2} = 0, \quad (\Delta \operatorname{curl} a)_\tau = (P\Delta v_1)_\tau.$$

Here P is the Hodge projection, and the subscript τ denotes the components tangential to $\partial\Omega$. The existence of an a satisfying (19) follows from the trace theorem [10]. We now construct v in three parts: $v = v_0(x) + \hat{v}(x,t) + v^*(x,t)$. We make $\hat{v}$ satisfy the initial conditions

$$(20) \qquad \begin{aligned} &\hat{v}(x,0) = 0, \quad \dot{\hat{v}}(x,0) = v_1(x) - \operatorname{curl} a(x), \\ &\ddot{\hat{v}}(x,0) = v_2(x), \quad \frac{\partial^3}{\partial t^3}\hat{v}(x,0) = v_3(x). \end{aligned}$$

With $-A = P\Delta$ denoting the Stokes operator, we have $v_3 \in D(A^{1/2})$, $v_2 \in D(A)$, and $v_1 - \operatorname{curl} a \in D(A^{3/2})$. The trace theorem [10] yields the existence of $\hat{v}$ satisfying (20) and

$$(21) \qquad \hat{v} \in \bigcap_{k=0}^{4} C^k([0,T'];D(A^{2-k/2})).$$

This implies in particular that $\hat{v}$ satisfies $(17)_1$, $(17)_3$, and $(17)_4$. Next, we construct v^* as curl a^*, where a^* is required to satisfy the initial conditions

$$(22) \qquad a^*(x,0) = 0, \quad \dot{a}^*(x,0) = a(x), \quad \ddot{a}^*(x,0) = 0, \quad \frac{\partial^3}{\partial t^3}a^*(x,0) = 0.$$

It follows from (19) that $a \in D(B)$, where B is the biharmonic operator with Dirichlet boundary conditions. The trace theorem implies the existence of an a^* satisfying (22) and

$$(23) \qquad a^* \in \bigcap_{k=0}^{5} C^k([0,T'];D(B^{(5-k)/4})).$$

This implies that v^* satisfies $(17)_1$, $(17)_3$, and $(17)_4$. □

To show that Σ is a contraction on $Z(M,T')$, we need appropriate estimates for the solutions of (15) and (16). These estimates will involve bounds of the form $\psi(M,T',\alpha,\beta,\cdots)$, where we may be compelled to choose M large, while α, β, $\cdots$ can be kept within prescribed bounds. It is then important that the size of ψ for large M can be controlled by choosing T' small. This motivates the following definition.

DEFINITION. A continuous function $\psi(M,T',\alpha,\beta,\cdots) : \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \cdots \to \mathbb{R}^+$ is called controllable if there are continuous functions $\tau(M,\alpha,\beta,\cdots)$ and $\omega(\alpha,\beta,\cdots)$ such that $\tau > 0$ and $\psi(M,T',\alpha,\beta,\cdots) \leq \omega(\alpha,\beta,\cdots)$ as long as $T' \leq \tau(M,\alpha,\beta,\cdots)$.

In the following, we regard the body force f and the initial conditions as given, and hence the dependence of estimates on these quantities will be suppressed.

The solution of (15) is routine. Any of the methods commonly used to solve hyperbolic problems (characteristics, semigroups, Galerkin) is applicable. Note that after extending v^n and $\mathbf{T}_0$ to all of $\mathbb{R}^3$ we may regard (15) as an initial value problem posed on all of space. Without proof, we state the following lemma.

LEMMA 2. *Let $v^n \in Z(M,T')$ be given and assume that T' is sufficiently small relative to M. Then* (15) *has a unique solution*

$$\mathbf{T}^n \in \bigcap_{k=0}^{3} C^k([0,T'];H^{3-k}(\Omega)).$$

Moreover, $\mathbf{T}^n$ satisfies a bound of the form

$$\|\mathbf{T}^n\|_{0,3} + \|\mathbf{T}^n\|_{1,2} + \|\mathbf{T}^n\|_{2,1} + \|\mathbf{T}^n\|_{3,0} \leq \phi(M,T'), \tag{24}$$

where ϕ is controllable. If v, $\hat{v}$ are two functions in $Z(M,T')$, then the corresponding solutions of (15) *satisfy an estimate of the form*

$$\begin{aligned} &\|\mathbf{T}-\hat{\mathbf{T}}\|_{0,2} + \|\mathbf{T}-\hat{\mathbf{T}}\|_{1,1} + \|\mathbf{T}-\hat{\mathbf{T}}\|_{2,0} \\ &\qquad \leq \psi(M,T')[\|v-\hat{v}\|_{0,3} + \|v-\hat{v}\|_{1,2} + \|v-\hat{v}\|_{2,1} + \|v-\hat{v}\|_{3,0}]. \end{aligned} \tag{25}$$

For any M, the function $\psi(M,T')$ tends to zero as $T' \to 0$.

We now turn to equation (16). To simplify notation, we suppress the index n; we write v for v^n and w for v^{n+1}. Moreover, d/dt shall denote $\partial/\partial t + (v \cdot \nabla)$. Problem (16) is then of the form

$$\begin{aligned} &\rho\frac{d^2w_i}{dt^2} = -\frac{\partial q}{\partial x_i} + C_{ijkl}(x,t)\frac{\partial^2 w_k}{\partial x_j \partial x_l} + h_i, \\ &\text{div } w = 0, \quad w|_{\partial\Omega} = 0, \quad w(x,0) = v_0(x), \quad \dot{w}(x,0) = v_1(x). \end{aligned} \tag{26}$$

LEMMA 3. *Let v, h, and $\mathbf{C}$ be given such that* div $v = 0$, $v|_{\partial\Omega} = 0$ *and $\mathbf{C}$ satisfies* (10), (11) *with a positive lower bound γ for κ in* (11). *Moreover, assume that bounds of the following kind hold:*

$$\|v\|_{0,4} + \|v\|_{1,3} + \|v\|_{2,2} + \|v\|_{3,1} + \|v\|_{4,0} \leq M, \tag{27}_1$$

$$\|v\|_{0,3} + \|v\|_{1,2} + \|v\|_{2,1} + \|v\|_{3,0} \leq K, \tag{27}_2$$

$$(27)_3 \qquad \|\mathbf{C}\|_{0,3} + \|\mathbf{C}\|_{1,2} + \|\mathbf{C}\|_{2,1} + \|\mathbf{C}\|_{3,0} \leq M,$$

$$(27)_4 \qquad \|\mathbf{C}\|_{0,2} + \|\mathbf{C}\|_{1,1} + \|\mathbf{C}\|_{2,0} \leq K,$$

$$(27)_5 \qquad \|h\|_{0,2} + \|h\|_{1,1} + \|h\|_{2,0} \leq K,$$

$$(27)_6 \qquad \left\|\frac{dh}{dt}\right\|_{0,2,1} + \left\|\frac{dh}{dt}\right\|_{1,1,1} + \left\|\frac{dh}{dt}\right\|_{2,0,1} \leq K.$$

Finally, assume that the initial values of $\ddot{w}$ *and* $\partial^3 w/\partial t^3$, *as determined by* (26), *are compatible with the boundary condition* $w|_{\partial\Omega} = 0$. *Then* (26) *has a solution*

$$w \in \bigcap_{k=0}^{4} C^k([0,T']; H^{4-k}(\Omega)),$$

and we have

$$(28) \qquad \|w\|_{0,4} + \|w\|_{1,3} + \|w\|_{2,2} + \|w\|_{3,1} + \|w\|_{4,0} \leq \phi(M,T',K,\gamma),$$

where ϕ *is controllable.*

It follows in a straightforward manner from Lemmas 2 and 3 that Σ maps $Z(M,T')$ into itself if we choose M large enough and T' small enough. Note that in evaluating $[\partial/\partial t + (v^n \cdot \nabla)]h(v^n, \nabla v^n, \dot{v}^n, \mathbf{T}^n, \nabla\mathbf{T}^n, f, \nabla f, \dot{f})$, we may substitute the right-hand side of (15) for $[\partial/\partial t + (v^n \cdot \nabla)]\mathbf{T}^n$. In order to establish that Σ is a contraction, we need the next lemma.

LEMMA 4. *Consider* (26) *and a second equation*

$$(29) \qquad \begin{aligned} &\rho\Big[\frac{\partial}{\partial t} + (\tilde{v} \cdot \nabla)\Big]^2 \tilde{w}_i = -\frac{\partial \tilde{q}}{\partial x_i} + \tilde{C}_{ijkl}(x,t)\frac{\partial^2 \tilde{w}_k}{\partial x_j \partial x_l} + \tilde{h}_i, \\ &\operatorname{div} \tilde{w} = 0, \quad \tilde{w}|_{\partial\Omega} = 0, \quad \tilde{w}(x,0) = v_0(x), \quad \dot{\tilde{w}}(x,0) = v_1(x). \end{aligned}$$

Assume that the assumptions of Lemma 3 *also hold for* (29) *(with the same constants* M, K *and* γ*). Moreover, assume that for* $t = 0$ *we have*

$$(30) \qquad v = \tilde{v}, \quad \dot{v} = \dot{\tilde{v}}, \quad \ddot{v} = \ddot{\tilde{v}}, \quad \mathbf{C} = \tilde{\mathbf{C}}, \quad \dot{\mathbf{C}} = \dot{\tilde{\mathbf{C}}}, \quad h = \tilde{h}, \quad \dot{h} = \dot{\tilde{h}}.$$

Then we have an estimate of the form

$$\begin{aligned} &\|w - \tilde{w}\|_{0,3} + \|w - \tilde{w}\|_{1,2} + \|w - \tilde{w}\|_{2,1} + \|w - \tilde{w}\|_{3,0} \\ &\quad \leq \psi(M,T',K,\gamma)\Big[\|v - \tilde{v}\|_{0,3} + \|v - \tilde{v}\|_{1,2} + \|v - \tilde{v}\|_{2,1} + \|v - \tilde{v}\|_{3,0} \\ &\qquad + \|\mathbf{C} - \tilde{\mathbf{C}}\|_{0,2} + \|\mathbf{C} - \tilde{\mathbf{C}}\|_{1,1} + \|\mathbf{C} - \tilde{\mathbf{C}}\|_{2,0} \\ &\qquad + \|h - \tilde{h}\|_{0,1} + \|h - \tilde{h}\|_{1,0} \\ &\qquad + \Big\|\frac{\partial h}{\partial t} + (v \cdot \nabla)h - \frac{\partial \tilde{h}}{\partial t} - (\tilde{v} \cdot \nabla)\tilde{h}\Big\|_{0,1,1} \\ &\qquad + \Big\|\frac{\partial h}{\partial t} + (v \cdot \nabla)h - \frac{\partial \tilde{h}}{\partial t} - (\tilde{v} \cdot \nabla)\tilde{h}\Big\|_{1,0,1}\Big]. \end{aligned} \qquad (31)$$

The function $\psi(M, T', K, \gamma)$ tends to zero as $T' \to 0$.

It is now an immediate consequence of Lemmas 2 and 4 that Σ is a contraction if T' is chosen small enough.

Even though (26) is a linear equation, the scheme we shall use to construct solutions is quite involved. The Galerkin approach (with divergence-free basis functions) is naturally suited to problems like (26); however, we seek solutions of a fairly high level of regularity, and we shall therefore use the Galerkin method not on (26) itself, but on an equation obtained after taking several derivatives of (26). To this purpose, we first define an iteration scheme, which will be described in detail in §3. We define a new variable z, which is essentially d^2w/dt^2, and take two material time derivatives of (26). This will lead to a "hyperbolic" second-order equation for z, which is of the same form as (26), and an elliptic equation arising from the original equation (26), by which w can be reconstructed from z. The iteration is constructed by alternating between these two problems. A technical complication arises, because in some terms involving $\dot{w}$, which appear in the equation resulting from differentiating the incompressibility condition, we need an approximation which has a higher temporal regularity than the iterates $\dot{w}^n$. This is achieved by taking $\dot{w}^n$ and a second approximation to $\dot{w}$, obtained essentially by time integration of z^n, and applying to the pair a projection operator from a suitable pair of Hilbert spaces onto the diagonal. This yields a new approximation to $\dot{w}$, which combines the regularity properties of both the original ones.

In §4, we discuss the solution of the "hyperbolic" part of the iteration from §3. We use the Galerkin method, but we first take one more material time derivative of the equation. That is, the variable approximated by the Galerkin scheme is not z but dz/dt, and we shall have to introduce an auxiliary elliptic problem to determine z itself. Once the construction of solutions to equation (26) is accomplished, the proof of Lemmas 3 and 4 follows rather easily as a consequence of the construction scheme. The proof of the lemmas is completed at the end of §4.

3. Iterative solution of equation (26). We shall now define a procedure to construct solutions of equation (26). We make the assumptions of Lemma 3, and M, T', K, and γ have the same meanings which they have there. In (26), we set $z = d^2w/dt^2 - \lambda w$, where λ is a positive constant to be chosen below. We can then write (26) in the form

$$\begin{aligned} \rho z_i &= -\frac{\partial q}{\partial x_i} + C_{ijkl}(x,t)\frac{\partial^2 w_k}{\partial x_j \partial x_l} - \lambda \rho w_i + h_i, \\ &\operatorname{div} w = 0, \quad w|_{\partial\Omega} = 0. \end{aligned} \tag{32}$$

We shall regard (32) as an elliptic problem from which we determine w for a given z. An equation for z is obtained by applying the operator $d/dt + (\nabla v)^T$ twice to equation (26). By doing this, we find an equation of the form

$$\begin{aligned} \rho\frac{d^2 z_i}{dt^2} = &-\frac{\partial \phi}{\partial x_i} + C_{ijkl}\frac{\partial^2 z_k}{\partial x_j \partial x_l} - \lambda \rho z_i \\ &+ G_i(z, \nabla z, \dot{z}, w, \nabla w, \nabla^2 w, \nabla^3 w, \dot{w}, \nabla \dot{w}, \nabla^2 \dot{w}, v, \nabla v, \nabla^2 v, \dot{v}, \nabla \dot{v}, \\ &\quad h, \nabla h, \nabla^2 h, \dot{h}, \nabla \dot{h}, \ddot{h}, \mathbf{C}, \nabla \mathbf{C}, \nabla^2 \mathbf{C}, \dot{\mathbf{C}}, \nabla \dot{\mathbf{C}}, \ddot{\mathbf{C}}). \end{aligned} \tag{33}$$

Here we have set $\phi := d^2q/dt^2$. From (26), we can find initial conditions for z and $\dot{z}$,

$$z(x,0) = z_0(x), \qquad \dot{z}(x,0) = z_1(x), \tag{34}$$

where $\|z_0\|_2 + \|z_1\|_1$ can be bounded in terms of K and λ. The boundary condition for z is

$$z = 0 \quad \text{on } \partial\Omega. \tag{35}$$

Moreover, after some algebra, we find

$$\begin{aligned} \operatorname{div} z &= \operatorname{div} \left[2(\dot{w}\cdot\nabla)v + (\dot{v}\cdot\nabla)w + \{[(v\cdot\nabla)w]\cdot\nabla\}v + v \operatorname{div}[(w\cdot\nabla)v]\right] \\ &=: \operatorname{div} H(v, \nabla v, \dot{v}, \nabla w, \dot{w}). \end{aligned} \tag{36}$$

We apply the operation d/dt to (36) and obtain an equation of the form

$$\begin{aligned} \operatorname{div} \frac{dz}{dt} &= \operatorname{div}\left[\frac{dH}{dt} + (z\cdot\nabla v) - (H\cdot\nabla)v\right] \\ &=: \operatorname{div} \tilde{H}(v, \nabla v, \nabla^2 v, \dot{v}, \nabla\dot{v}, \ddot{v}, w, \nabla w, \nabla^2 w, \dot{w}, \nabla\dot{w}, z). \end{aligned} \tag{37}$$

We note that H (and hence $\tilde{H}$) vanishes on $\partial\Omega$.

The solution is now constructed in an iterative fashion. With z^n given, we obtain w^n by solving the elliptic problem

$$\begin{aligned} \rho z_i^n &= -\frac{\partial q^n}{\partial x_i} + C_{ijkl}\frac{\partial^2 w_k^n}{\partial x_j \partial x_l} - \lambda\rho w_i^n + h_i, \\ \operatorname{div} w^n &= 0, \quad w^n|_{\partial\Omega} = 0. \end{aligned} \tag{38}$$

Then we find z^{n+1} from the problem

$$\begin{aligned} \rho\frac{d^2 z_i^{n+1}}{dt^2} &= -\frac{\partial \phi^{n+1}}{\partial x_i} + C_{ijkl}\frac{\partial^2 z_k^{n+1}}{\partial x_j \partial x_l} - \lambda\rho z_i^{n+1} \\ &\quad + G_i(z^n, \nabla z^n, \dot{z}^n, w^n, \nabla w^n, \nabla^2 w^n, \nabla^3 w^n, \dot{w}^n, \nabla\dot{w}^n, \nabla^2\dot{w}^n, \\ &\qquad v, \nabla v, \nabla^2 v, \dot{v}, \nabla\dot{v}, h, \nabla h, \nabla^2 h, \dot{h}, \nabla\dot{h}, \ddot{h}, \mathbf{C}, \nabla\mathbf{C}, \nabla^2\mathbf{C}, \dot{\mathbf{C}}, \nabla\dot{\mathbf{C}}, \ddot{\mathbf{C}}), \\ \operatorname{div}\frac{dz^{n+1}}{dt} &= \operatorname{div}\tilde{H}(v, \nabla v, \nabla^2 v, \dot{v}, \nabla\dot{v}, \ddot{v}, w^n, \nabla w^n, \nabla^2 w^n, \psi^n, \nabla\psi^n, z^n), \\ z^{n+1}(x,0) &= z_0(x), \quad \dot{z}^{n+1}(x,0) = z_1(x), \quad z^{n+1}|_{\partial\Omega} = 0. \end{aligned} \tag{39}$$

Here ψ^n is an approximation to $\dot{w}$ which is in general different from $\dot{w}^n$ and will be defined below. In order for the iteration to converge, we may have to restrict to a time interval $[0, T^*]$ with $T^* < T'$; since (26) is linear, we can always continue the solution.

We seek our solution in the function space $X(L, T^*)$, defined as the set of all functions z such that

$$z \in \bigcap_{k=0}^{2} W^{k,\infty}([0, T^*]; H^{2-k}(\Omega)), \tag{40}_1$$

$$\|z\|_{0,2} + \|z\|_{1,1} + \|z\|_{2,0} \le L, \tag{40}_2$$

$$z|_{\partial\Omega} = 0, \tag{40}_3$$

$$z(x,0) = z_0(x), \qquad \dot{z}(x,0) = z_1(x). \tag{40}_4$$

If L is chosen large enough, $X(L,T^*)$ is not empty. For given $z \in X(L,T^*)$, we can solve the problem (32).

LEMMA 5. *Let $\lambda > 0$ be sufficiently large; how large depends only on K and γ. Then, for every $z \in X(L,T^*)$, equation* (32) *has a unique solution w such that*

$$w \in \bigcap_{k=0}^{2} W^{k,\infty}([0,T^*]; H^{4-k}(\Omega)), \tag{41}_1$$

$$\|w\|_{0,4} + \|w\|_{1,3} + \|w\|_{2,2} \le \alpha(K,\gamma)L + \beta(K,\gamma), \tag{41}_2$$

$$w(x,0) = v_0(x), \qquad \dot{w}(x,0) = v_1(x). \tag{41}_3$$

The proof of Lemma 5 employs standard techniques of elliptic theory and no details are provided. First a weak solution (in $H^1(\Omega)$) is constructed using variational techniques. Following the lines of Giaquinta and Modica [5], it can be shown that the weak solution is actually in $H^2(\Omega)$. Higher regularity follows from the results of Agmon, Douglis, and Nirenberg [1].

We now describe the construction of ψ^n in (39). From the definition of z, we obtain

$$z = \ddot{w} + 2(v\cdot\nabla)\dot{w} + (\dot{v}\cdot\nabla)w + (v\cdot\nabla)(v\cdot\nabla)w - \lambda w, \tag{42}$$

or, after integration,

$$\begin{aligned}\dot{w}(x,t) = v_1(x) + \int_0^t z(x,\tau) - 2(v\cdot\nabla)\dot{w}(x,\tau) - (\dot{v}\cdot\nabla)w(x,\tau) \\ - (v\cdot\nabla)(v\cdot\nabla)w(x,\tau) + \lambda w(x,\tau)\, d\tau.\end{aligned} \tag{43}$$

However, the iterates z^n, w^n cannot be expected to satisfy this identity. If z^n is in $X(L,T^*)$, then Lemma 5 guarantees that $\dot{w}^n \in L^\infty([0,T^*]; H^3(\Omega)\cap H_0^1(\Omega)) \cap W^{1,\infty}([0,T^*]; H^2(\Omega)\cap H_0^1(\Omega))$, while the iterates corresponding to the expression on the right-hand side of (43) lie in $W^{1,\infty}([0,T^*]; H^2(\Omega)\cap H_0^1(\Omega))\cap W^{2,\infty}([0,T^*]; H_0^1(\Omega))$. Thus one of the two approximations for $\dot{w}$ has more spatial regularity, the other has more temporal regularity. The function ψ^n is chosen such that it combines both spatial and temporal regularity.

Let χ be a function such that

$$\chi \in \bigcap_{k=0}^{2} W^{k,\infty}([0,T^*]; H^{3-k}(\Omega)\cap H_0^1(\Omega)), \tag{44}_1$$

$$\chi(x,0) = v_1(x), \ \dot{\chi}(x,0) = w_2(x), \tag{44}_2$$

where

$$w_2 = z_0 - 2(v\cdot\nabla)v_1 - (\dot{v}\cdot\nabla)v_0 - (v\cdot\nabla)(v\cdot\nabla)v_0 + \lambda v_0 \tag{44}_3$$

is the initial value of $\ddot{w}$. We set

$$\begin{aligned}
&\alpha^n(x,t) = \dot{w}^n(x,t) - \chi(x,t),\\
&\beta^n(x,t) = v_1(x) + \int_0^t z^n(x,\tau) - 2(v\cdot\nabla)\dot{w}^n(x,\tau) - (\dot{v}\cdot\nabla)w^n(x,\tau)\\
&\qquad\qquad - (v\cdot\nabla)(v\cdot\nabla)w^n(x,\tau) + \lambda w^n(x,\tau)\,d\tau - \chi(x,t).
\end{aligned} \tag{45}$$

Let

$$\begin{aligned}
X &= \{u \in L^\infty([0,T^*]; H^3(\Omega)\cap H_0^1(\Omega))\\
&\qquad \cap W^{1,\infty}([0,T^*]; H^2(\Omega)\cap H_0^1(\Omega)) \mid u(x,0) = 0\},\\
Y &= \{u \in W^{1,\infty}([0,T^*]; H^2(\Omega)\cap H_0^1(\Omega))\\
&\qquad \cap W^{2,\infty}([0,T^*]; H_0^1(\Omega)) \mid u(x,0) = \dot{u}(x,0) = 0\},\\
Z &= \left\{u \in \bigcap_{k=0}^{2} H^k([0,T^*]; H^{3-k}(\Omega)\cap H_0^1(\Omega)) \mid u(x,0) = \dot{u}(x,0) = 0\right\}.
\end{aligned} \tag{46}$$

We shall construct a continuous linear mapping $\Pi : X \times Y \to Z$ with the following properties:

(i) $\Pi(u,u) = u$ for every $u \in X \cap Y$.

(ii) The norm of Π has a bound which is independent of T^* as $T^* \to 0$.

We define ψ^n as

$$\psi^n = \chi + \Pi(\alpha^n, \beta^n). \tag{47}$$

We now describe the construction of the operator Π. Let α be in either X or Y. Then we define a temporally periodic extension $E_1\alpha$ as follows:

$$\begin{aligned}
E_1\alpha(x,t) &= \begin{cases} \alpha(x,t) & \text{if } t \in [0,T^*],\\ 2\alpha(x,T^*) - \alpha(x,2T^*-t) & \text{if } t \in [T^*,2T^*],\\ E_1\alpha(x,-t) & \text{if } t \in [-2T^*,0]. \end{cases}\\
E_1\alpha(x,t+4T^*) &= E_1\alpha(x,t).
\end{aligned} \tag{48}$$

We note that the temporal average of $E_1\alpha$ is $\alpha(x,T^*)$. Let E_2 be an extension operator which maps $H^k(\Omega)$ to $H^k(\mathbb{R}^3)$ for $k = 1,2,3$. We can choose E_2 independent of k.

Let $H_p^k(\mathbb{R}; V)$ denote the space of all $4T^*$-periodic functions $\mathbb{R} \to V$ with H^k-regularity, and let P be the orthogonal projection from $L_p^2(\mathbb{R}; H^2(\mathbb{R}^3)) \times H_p^1(\mathbb{R}; H^1(\mathbb{R}^3))$ onto the diagonal $L_p^2(\mathbb{R}; H^2(\mathbb{R}^3)) \cap H_p^1(\mathbb{R}; H^1(\mathbb{R}^3))$. The following properties of P are easily checked:

(i) P is continuous from $(L_p^2(\mathbb{R}; H^3(\mathbb{R}^3)) \cap H_p^1(\mathbb{R}; H^2(\mathbb{R}^3))) \times (H_p^1(\mathbb{R}; H^2(\mathbb{R}^3)) \cap H_p^2(\mathbb{R}; H^1(\mathbb{R}^3)))$ onto $\bigcap_{k=0}^2 H_p^k(\mathbb{R}; H^{3-k}(\mathbb{R}^3))$.

(ii) If α and β are even functions of time, then so is $P(\alpha,\beta)$.

(iii) If α and β are constant (in time), then so is $P(\alpha,\beta)$.

(iv) If α and β have zero temporal average, then so does $P(\alpha,\beta)$.

Hence the temporal average of $P(\alpha,\beta)$ depends only on those of α and β.

Let Υ be the operator of evaluation at $t = 0$. By the trace theorem, Υ is continuous

$$\bigcap_{k=0}^{2} H_p^k(\mathbb{R}; H^{3-k}(\mathbb{R}^3)) \to H^{5/2}(\mathbb{R}^3). \tag{49}_1$$

Moreover, the restriction of Υ to functions of zero average has a norm which is bounded independently of T^* as $T^* \to 0$. Let E_3 be a right inverse of Υ which maps

$$(49)_2 \qquad H^{5/2}(\mathbb{R}^3) \to \{u \in H^3(\mathbb{R}^3 \times [0,T^*]) \mid \dot{u}(x,0) = 0\}.$$

Let R be the operator of restriction to $\Omega \times [0,T^*]$ and let $Q : H^1(\Omega) \to H_0^1(\Omega)$ be the solution operator ($f \mapsto u$) for the problem

$$\Delta u = \Delta f, \qquad u|_{\partial\Omega} = 0. \tag{50}$$

We define

$$\Pi(\alpha,\beta) = QR(Id - E_3\Upsilon)P(E_2E_1\alpha, E_2E_1\beta). \tag{51}$$

It is easy to verify that Π has the properties (i) and (ii) above.

For later use, we also note the following lemma.

LEMMA 6. *There is a constant C, independent of T^* as $T^* \to 0$, such that*

$$\begin{aligned} &\|\alpha - \Pi(\alpha,\beta)\|_{0,3,2} + \|\alpha - \Pi(\alpha,\beta)\|_{1,2,2} \\ &\quad + \|\beta - \Pi(\alpha,\beta)\|_{1,2,2} + \|\beta - \Pi(\alpha,\beta)\|_{2,1,2} \\ &\quad \le C\|\alpha - \beta\|_{1,2} \qquad \forall\ \alpha \in X,\ \beta \in Y. \end{aligned} \tag{52}$$

The proof follows easily from the correponding property of P: For $\alpha \in L_p^2(\mathbb{R}; H^3(\mathbb{R}^3)) \cap H_p^1(\mathbb{R}; H^2(\mathbb{R}^3))$ and $\beta \in H_p^1(\mathbb{R}; H^2(\mathbb{R}^3)) \cap H_p^2(\mathbb{R}; H^1(\mathbb{R}^3))$, we have

$$\begin{aligned} &\|\alpha - P(\alpha,\beta)\|_{0,3,2} + \|\alpha - P(\alpha,\beta)\|_{1,2,2} \\ &\quad + \|\beta - P(\alpha,\beta)\|_{1,2,2} + \|\beta - P(\alpha,\beta)\|_{2,1,2} \le C\|\alpha - \beta\|_{1,2,2}. \end{aligned} \tag{53}$$

This latter inequality can be verified using the explicit construction of P in terms of Fourier analysis. If

$$\begin{aligned} \alpha &= \sum_k e^{ik\pi t/2T^*} \int_{\mathbb{R}^3} \hat{\alpha}_k(\xi) e^{i\xi\cdot x}\, d\xi, \\ \beta &= \sum_k e^{ik\pi t/2T^*} \int_{\mathbb{R}^3} \hat{\beta}_k(\xi) e^{i\xi\cdot x}\, d\xi, \end{aligned} \tag{54}$$

then

$$P(\alpha,\beta) = \sum_k e^{ik\pi t/2T^*} \int_{\mathbb{R}^3} \hat{\gamma}_k(\xi) e^{i\xi\cdot x}\, d\xi, \tag{55}$$

where

$$\hat{\gamma}_k(\xi) = \frac{(1+|\xi|^2)\hat{\alpha}_k(\xi) + (1 + (k^2\pi^2/4T^{*2}))\hat{\beta}_k(\xi)}{2 + |\xi|^2 + (k^2\pi^2/4T^{*2})}. \tag{56}$$

4. Galerkin approximation. We now consider equation (39), which we restate as follows:

$$\begin{aligned} &\rho\frac{d^2z_i}{dt^2} = -\frac{\partial\phi}{\partial x_i} + C_{ijkl}\frac{\partial^2 z_k}{\partial x_j \partial x_l} - \lambda\rho z_i + G_i, \\ &\operatorname{div}\frac{dz}{dt} = \operatorname{div}\tilde{H}, \\ &z|_{\partial\Omega} = 0, \quad z(x,0) = z_0(x), \quad \dot{z}(x,0) = z_1(x). \end{aligned} \tag{57}$$

We make the following assumptions:

(s1)
$$v \in \bigcap_{k=0}^{3} W^{k,\infty}([0,T^*];H^{3-k}(\Omega)),$$
$$\|v\|_{3,0} + \|v\|_{2,1} + \|v\|_{1,2} + \|v\|_{0,3} \le K.$$

(s2)
$$\mathbf{C} \in \bigcap_{k=0}^{3} W^{k,\infty}([0,T^*];H^{3-k}(\Omega)),$$
$$\|\mathbf{C}\|_{3,0} + \|\mathbf{C}\|_{2,1} + \|\mathbf{C}\|_{1,2} + \|\mathbf{C}\|_{0,3} \le M,$$
$$\|\mathbf{C}\|_{2,0} + \|\mathbf{C}\|_{1,1} + \|\mathbf{C}\|_{0,2} \le K.$$

(s3)
$$G \in C([0,T^*];L^2(\Omega)), \qquad \frac{dG}{dt} \in L^1([0,T^*];L^2(\Omega)),$$
$$\|G\|_{0,0} + \left\|\frac{dG}{dt}\right\|_{0,0,1} \le K.$$

(s4)
$$\tilde{H} \in \bigcap_{k=0}^{2} W^{k,1}([0,T^*];H^{2-k}(\Omega)),$$
$$\|\tilde{H}\|_{0,1} + \|\tilde{H}\|_{1,0} + \|\tilde{H}\|_{0,2,1} + \|\tilde{H}\|_{1,1,1} + \|\tilde{H}\|_{2,0,1} \le K.$$

(s5) $z_0 \in H^2(\Omega)$, $z_1 \in H^1(\Omega)$; since the initial data never change during the iteration procedure, the dependence on their norms will not be indicated in the estimates.

(e) $\mathbf{C}$ satisfies the symmetry condition (10) and strong ellipticity condition (11), with κ bounded from below by γ.

(c1) z_0 and z_1 vanish on $\partial\Omega$.

(c2) $\operatorname{div}(z_1 + (v\cdot\nabla)z_0) = \operatorname{div}\tilde{H}(\cdot,0)$.

(c3) $\tilde{H}$ vanishes on $\partial\Omega$.

LEMMA 7. *Assume that* (s1)–(s5), (e), *and* (c1)–(c3) *hold. Then* (57) *has a unique solution* $z \in \bigcap_{k=0}^{2} C^k([0,T^*];H^{2-k}(\Omega))$, *and* $\|z\|_{2,0}+\|z\|_{1,1}+\|z\|_{0,2} \le \sigma(M,T^*,K,\gamma)$, *where* σ *is controllable.*

Proof. Without loss of generality, we shall assume $\tilde{H} = 0$; with u denoting the solution of the problem

$$\frac{du}{dt} = \tilde{H}, \qquad u(x,0) = 0, \tag{58}$$

we may consider $z - u$ instead of z. We apply the operator $d/dt + (\nabla v)^T$ to (57), set $dz/dt = y$, $d\phi/dt = \omega$, and obtain

$$\begin{aligned}\rho\frac{d^2y_i}{dt^2} + \rho\frac{\partial v_m}{\partial x_i}\frac{dy_m}{dt} &= -\frac{\partial\omega}{\partial x_i} + C_{ijkl}\frac{\partial^2 y_k}{\partial x_j\partial x_l} - \lambda\rho y_i + \frac{dC_{ijkl}}{dt}\frac{\partial^2 z_k}{\partial x_j\partial x_l}\\ &\quad - C_{ijkl}\left[\frac{\partial^2 v_m}{\partial x_j\partial x_l}\frac{\partial z_k}{\partial x_m} + \frac{\partial v_m}{\partial x_j}\frac{\partial^2 z_k}{\partial x_l\partial x_m} + \frac{\partial v_m}{\partial x_l}\frac{\partial^2 z_k}{\partial x_j\partial x_m}\right]\\ &\quad + \frac{\partial v_m}{\partial x_i}C_{mjkl}\frac{\partial^2 z_k}{\partial x_j\partial x_l} - \lambda\rho\frac{\partial v_m}{\partial x_i}z_i + \frac{dG_i}{dt} + \frac{\partial v_m}{\partial x_i}G_m,\end{aligned} \tag{59}$$

$$\operatorname{div} y = 0, \quad y|_{\partial\Omega} = 0, \quad y(x,0) = \tilde{z}_1(x) := z_1(x) + (v\cdot\nabla)z_0(x),$$
$$\frac{dy}{dt}(x,0) = \tilde{z}_2(x),$$

where $\tilde{z}_2$ is the appropriate initial value determined from (57).

Our task is now to solve (59) for y, where we think of z as being determined in terms of y by the elliptic problem

$$\begin{aligned} &\rho\frac{dy_i}{dt} = -\frac{\partial\phi}{\partial x_i} + C_{ijkl}\frac{\partial^2 z_k}{\partial x_j \partial x_l} - \lambda\rho z_i + G_i, \\ &\frac{d}{dt}\text{div } z = -\text{div }((z\cdot\nabla)v), \quad \text{div } z(\cdot,0) = \text{div } z_0, \quad z|_{\partial\Omega} = 0. \end{aligned} \tag{60}$$

This takes a proof that by solving (59) and (60) we really obtain a solution of (57).

PROPOSITION 8. *Let* $y \in \bigcap_{k=0}^{1} W^{k,\infty}([0,T^*]; H^{1-k}(\Omega))$ *and* $z \in L^\infty([0,T^*]; H^2(\Omega)) \cap W^{1,1}([0,T^*]; H^1(\Omega))$ *be such that* (59) *and* (60) *hold. Then* $dz/dt = y$.

Proof. We apply the operator $d/dt + (\nabla v)^T$ to equation (60). By comparing with (59), we find that $\eta := dz/dt - y$ satisfies the system

$$\begin{aligned} &0 = -\frac{\partial}{\partial x_i}\left(\frac{d\phi}{dt} - \omega\right) + C_{ijkl}\frac{\partial^2 \eta_k}{\partial x_j \partial x_l} - \lambda\rho\eta_i, \\ &\text{div } \eta = 0, \qquad \eta|_{\partial\Omega} = 0. \end{aligned} \tag{61}$$

It is immediate from this that $\eta = 0$. □

To solve (59), (60), we use a Galerkin approximation. Let V be the space of all divergence-free vector fields in $H_0^1(\Omega)$, and let $\{\tilde{\phi}^i | i \in \mathbb{N}\}$ be a basis for V. We define a time-dependent basis of V as follows:

$$\frac{d\phi^i}{dt} - (\phi^i\cdot\nabla)v = 0, \qquad \phi^i(x,0) = \tilde{\phi}^i(x). \tag{62}$$

We seek an approximation to y of the form

$$y^N(x,t) = \sum_{n=1}^{N} \alpha^n(t)\phi^n(x,t). \tag{63}$$

For given y^N, let z^N be the corresponding solution of (60). We solve the following approximate version of (59). For $n = 1, 2, \cdots, N$, we require that

$$\begin{aligned} \int_\Omega \phi_i^n \Bigg\{ &\rho\frac{d^2 y_i^N}{dt^2} + \rho\frac{\partial v_m}{\partial x_i}\frac{dy_m^N}{dt} - C_{ijkl}\frac{\partial^2 y_k^N}{\partial x_j \partial x_l} + \lambda\rho y_i^N \\ &- \frac{dC_{ijkl}}{dt}\frac{\partial^2 z_k^N}{\partial x_j \partial x_l} + C_{ijkl}\left[\frac{\partial^2 v_m}{\partial x_j \partial x_l}\frac{\partial z_k^N}{\partial x_m} + \frac{\partial v_m}{\partial x_j}\frac{\partial^2 z_k^N}{\partial x_l \partial x_m} + \frac{\partial v_m}{\partial x_l}\frac{\partial^2 z_k^N}{\partial x_j \partial x_m}\right] \\ &- \frac{\partial v_m}{\partial x_i}C_{mjkl}\frac{\partial^2 z_k^N}{\partial x_j \partial x_l} + \lambda\rho\frac{\partial v_m}{\partial x_i}z_i^N - \frac{dG_i}{dt} - \frac{\partial v_m}{\partial x_i}G_m \Bigg\} dx = 0, \end{aligned} \tag{64}$$

and we impose the initial conditions

$$\begin{aligned} &y^N(x,0) = P_1^N \tilde{z}_1(x), \\ &\frac{dy^N}{dt}(x,0) - (y^N(x,0)\cdot\nabla)v(x,0) = P_0^N(\tilde{z}_2 - (\tilde{z}_1\cdot\nabla)v)(x,0). \end{aligned} \tag{65}$$

Here P_1^N is the orthogonal projection in V onto the span of $\tilde{\phi}^1, \tilde{\phi}^2, \cdots, \tilde{\phi}^N$, and P_0^N is the orthogonal projection in $L^2(\Omega)$. We note that the combination $dy^N/dt - (y^N\cdot\nabla)v$

is equal to $\sum_{n=1}^{N} \dot{\alpha}^n(t)\phi^n(x,t)$, and hence (65) prescribes initial conditions for α^n and $\dot{\alpha}^n$. The problem (64), (65) is therefore an initial value problem for a linear system of second-order ODE's, and existence and uniqueness of solutions are trivial.

To obtain an estimate uniform in N, we observe that ϕ^n in (64) can be replaced by any linear combination of the ϕ^n, and we choose the special linear combination $dy^N/dt - (y^N \cdot \nabla)v$. We then integrate with respect to time. This yields, after some integrations by parts, an energy equation of the form

$$
\begin{aligned}
E_N(t) &:= \frac{1}{2}\int_\Omega \rho\left|\frac{dy^N}{dt}\right|^2 + C_{ijkl}\frac{\partial y_k^N}{\partial x_l}\frac{\partial y_i^N}{\partial x_j} + \lambda\rho|y^N|^2\,dx \\
&= E_N(0) + \int_0^t \int_\Omega \cdots\,dx\,dt.
\end{aligned} \tag{66}
$$

From this energy equation, we find uniform bounds for the norm of y^N in $L^\infty([0,T^*]; V) \cap W^{1,\infty}([0,T^*]; L^2(\Omega))$. We can extract a weakly-$*$ convergent subsequence, and it is straightforward to show that the limit is a solution of (59), (60) (cf. [10, p. 268] for a similar argument). It also follows from (59) that $y \in W^{2,1}([0,T^*]; V')$, from (60), we get $z \in W^{1,1}([0,T^*]; H_0^1(\Omega))$, and hence Proposition 8 is applicable. Altogether, this proves the existence of a solution to (57), which lies in $\bigcap_{k=0}^{2} W^{k,\infty}([0,T^*]; H^{2-k}(\Omega))$. Uniqueness follows from a straightforward estimate; multiply (57) by dz/dt and integrate with respect to space and time (we omit the details). The bound on the norm claimed in Lemma 7 also follows from the energy estimate (66); the controllability of the function σ arises from the fact that derivatives of $\mathbf{C}$ appear only under the time integral in (66).

It remains to be shown that we actually have $z \in \bigcap_{k=0}^{2} C^k([0,T^*]; H^{2-k}(\Omega))$. From the regularity already established, it follows that $z \in \bigcap_{k=0}^{2} C_w^k([0,T^*]; H^{2-k}(\Omega))$ (weak continuity). The strong continuity then follows if we can show that the energy function $E(t)$, defined as in (66), is continuous (see [14]). By considering the limit $N \to \infty$ in (66), we find that

$$
\limsup_{t\to 0+} E(t) \le E(0), \tag{67}
$$

while weak continuity implies

$$
E(0) \le \liminf_{t\to 0+} E(t). \tag{68}
$$

Hence E is continuous from the right at $t = 0$. There is nothing particular about $t = 0$; we may just as well impose initial conditions at any other time and use the same method as above to construct the solution from these new initial data. Therefore E is continuous from the right for all t. Since equation (57) can also be solved backward in time, E is also continuous from the left. □

From Lemmas 5 and 7, it is easy to conclude that, for appropriately chosen L and T^*, the mapping $\Phi : z^n \mapsto z^{n+1}$ defined by (38), (39) maps $X(L,T^*)$ into itself. It is also easy to see that Φ is a contraction if T^* is chosen small enough. Let z be the limit of the iteration and let w be related to z by (38). In order to complete the proof of Lemma 3, it remains to be shown that we really have $z = d^2w/dt^2 - \lambda w$. Let η denote the difference $\eta := z - d^2w/dt^2 + \lambda w$. We apply the operation $d/dt + (\nabla v)^T$ twice to equation (38) and compare with (39). This yields

$$
0 = C_{ijkl}\frac{\partial^2 \eta_k}{\partial x_j \partial x_l} - \lambda\rho\eta_i - \frac{\partial}{\partial x_i}\left(\phi - \frac{d^2 q}{dt^2}\right). \tag{69}
$$

By retracing the steps leading to (36), (37), and comparing with the second equation of (39), we find

$$
\begin{aligned}
\frac{d}{dt}\ \mathrm{div}\ \eta = \mathrm{div}\ \Big\{ & 2(\eta\cdot\nabla)v - 3\Big\{[(v\cdot\nabla)(\psi-\dot w)]\cdot\nabla\Big\}v + 2((\psi-\dot w)\nabla\dot v) \\
& + (\dot v\cdot\nabla)(\psi-\dot w) + v\ \mathrm{div}\ [(\psi-\dot w)\cdot\nabla]v \\
& + 2(v\cdot\nabla)[(\psi-\dot w)\nabla]v - 2\Big\{[((\psi-\dot w)\cdot\nabla)v]\cdot\nabla\Big\}v\Big\}, \\
\mathrm{div}\ \eta(x,0) &= 0.
\end{aligned} \tag{70}
$$

Finally, it is clear that $\eta = 0$ on $\partial\Omega$. Using Lemma 6, we can estimate $\|\psi-\dot w\|_{0,3,2} + \|\psi-\dot w\|_{1,2,2}$ in terms of $\|\alpha-\beta\|_{1,2}$, where, according to (45),

$$
\alpha(x,t) - \beta(x,t) = \int_0^t \eta(x,\tau)\ d\tau. \tag{71}
$$

After integrating (70) with respect to time and using this estimate, it is clear that $\eta = 0$.

To prove Lemma 4, we subtract (29) from (26), which results in

$$
\begin{aligned}
&\rho\Big(\frac{\partial}{\partial t} + (v\cdot\nabla)\Big)^2(w_i-\tilde w_i) \\
&\quad = -\frac{\partial}{\partial x_i}(q-\tilde q) + C_{ijkl}\frac{\partial^2}{\partial x_j\partial x_l}(w_k-\tilde w_k) + (C_{ijkl}-\tilde C_{ijkl})\frac{\partial^2\tilde w_k}{\partial x_j\partial x_l} \\
&\qquad + \rho\Big[\Big(\frac{\partial}{\partial t} + (v\cdot\nabla)\Big)^2 - \Big(\frac{\partial}{\partial t} + (\tilde v\cdot\nabla)\Big)^2\Big]\tilde w_i + h_i - \tilde h_i, \\
&\mathrm{div}\ (w-\tilde w) = 0, \qquad (w-\tilde w)|_{\partial\Omega} = 0, \\
&w(x,0) - \tilde w(x,0) = 0, \qquad \dot w(x,0) - \dot{\tilde w}(x,0) = 0.
\end{aligned} \tag{72}
$$

We apply the operation $\partial/\partial t + (v\cdot\nabla) + (\nabla v)^T$ twice to the first equation of (72), multiply by $L^3(w-\tilde w)$, where L is the operator $w \mapsto \partial w/\partial t + (v\cdot\nabla)w - (w\cdot\nabla)v$, and integrate with respect to space and time. Lemma 4 then follows from the resulting energy estimate. We omit the details of the calculation.

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] C. Chen and W. von Wahl, *Das Rand-Anfangswertproblem für quasilineare Wellengleichungen in Sobolevräumen niedriger Ordnung,* J. Reine Angew. Math., 337 (1982), pp. 77–112.

[3] C. M. Dafermos and W. J. Hrusa, *Energy methods for quasilinear hyperbolic initial-boundary value problems. Applications to elastodynamics,* Arch. Rational Mech. Anal., 87 (1985), pp. 267–292.

[4] E. G. Ebin and R. A. Saxton, *The initial value problem for elastodynamics of incompressible bodies,* Arch. Rational Mech. Anal., 94 (1986), pp. 15–38.

[5] M. Giaquinta and G. Modica, *Non-linear systems of the type of the stationary Navier-Stokes system,* J. Reine Angew. Math., 330 (1982), pp. 173–214.

[6] W. J. Hrusa and M. Renardy, *An existence theorem for the Dirichlet problem in the elastodynamics of incompressible materials,* Arch. Rational Mech. Anal., 102 (1988), pp. 95–117.

[7] T. J. R. HUGHES, T. KATO AND J. E. MARSDEN, *Well-posed quasi-linear second-order hyperbolic systems with applications to nonlinear elastodynamics and general relativity,* Arch. Rational Mech. Anal., 63 (1976), pp. 273–284.

[8] M. W. JOHNSON AND D. SEGALMAN, *A model for viscoelastic fluid behavior which allows non-affine deformation,* J. Non-Newt. Fluid Mech., 2 (1977), pp. 255–270.

[9] T. KATO, *Linear and quasi-linear equations of hyperbolic type,* in Hyperbolicity, G. da Prato and G. Geymonat, eds., Centro Internazionale Matematico Estivo, II ciclo, Cortona, 1976, pp. 125–191.

[10] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications* I, Springer-Verlag, Berlin, New York, 1972.

[11] J. G. OLDROYD, *Non-Newtonian effects in steady motion of some idealized elastico-viscous liquids,* Proc. Roy. Soc. London, Ser. A, 245 (1958), pp. 278–297.

[12] M. RENARDY, *Inflow boundary conditions for steady flows of viscoelastic fluids with differential constitutive laws,* Rocky Mtn. J. Math., 18 (1988), pp. 445–453.

[13] S. SCHOCHET, *The incompressible limit in nonlinear elasticity,* Comm. Math. Phys., 102 (1985), pp. 207–215.

[14] W. STRAUSS, *On continuity of functions with values in various Banach spaces,* Pacific J. Math., 19 (1966), pp. 543–551.

[15] R. TEMAM, *Navier-Stokes Equations,* Third edition, North-Holland, Amsterdam, 1984.

[16] C. A. TRUESDELL AND W. NOLL, *The non-linear field theories of mechanics,* in Handbuch der Physik III/3, S. Flügge, ed., Springer-Verlag, Berlin, New York, 1965.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1386-1393, November 1990
© 1990 Society for Industrial and Applied Mathematics

TRAVELING WAVE SOLUTIONS OF THE STEFAN AND THE ABLATION PROBLEMS*

RICCARDO RICCI†

Abstract. The problem of the stability of traveling wave solutions of the ablation and the Stefan problems is considered in the presence of convection. It is proved that under reasonable assumptions on the initial datum the solution has a traveling wave as an asymptotic limit, and the phase shift is computed.

Key words. Stefan problem, traveling waves, asymptotic behavior

AMS(MOS) subject classification. 35R35

1. Introduction. In this paper we consider the melting of a solid under the effect of an energy flux H. We suppose that the liquid phase is immediately removed from the vicinity of the melting front, for instance, by vaporization. This process, which is also known as ablation [4], [2], can be modeled as a one-phase Stefan problem modifying the heat balance equation on the melting front to take into account the heat flux entering the solid. In one space dimension, assuming that the solid occupies the space at the right of the free boundary $x = s(t)$ and indicating by $T(x, t)$ the temperature, we have

$$c\rho \frac{\partial T}{\partial t} - \frac{\partial}{\partial x}\left(\kappa \frac{\partial T}{\partial x}\right) = 0 \quad \text{in } x > s(t), \quad t > 0, \tag{1.1}$$

$$\kappa \frac{\partial T}{\partial x}(s(t), t) = \rho L \dot{s}(t) - H, \qquad t > 0. \tag{1.2}$$

The second equation is the energy balance at the free boundary and expresses the fact that the heat input H equals the rate of heat flow into the solid plus the rate of heat absorption by the melting. This condition can be seen to hold whether or not the solid actually melts. In the first case we assume thermal equilibrium at the melting front so that

$$T(s(t), t) = T_m \quad \text{where } \frac{ds}{dt}(t) > 0, \tag{1.3}$$

where T_m indicates the melting temperature.

If the temperature at $x = s(t)$ is less than the melting temperature, then the front does not move and (1.3) reads simply

$$\frac{ds}{dt}(t) = 0 \quad \text{where } T(s(t), t) < T_m. \tag{1.3'}$$

In this case (1.2) is just a Neumann condition for the heat equation in $x > s$.

A first detailed investigation of the ablation problem can be found in an old paper by Landau [8]. In this work the author considers both the case of a semi-infinite solid and that of a slab of finite thickness. No mathematical proof of existence or uniqueness of the solution of this problem is given but some useful integral relations between T and s are established, and some special solutions are given explicitly. In particular,

* Received by the editors December 5, 1988; accepted for publication (in revised form) December 6, 1989.
† Istituto di Matematica, Facoltà di Ingegneria dell'Università di Ancona, Via Brecce Bianche, 60100 Ancona, Italy.

Landau shows that the problem has a traveling wave solution. This is recovered immediately and its expression is

$$T(x,t) = T_0 + (T_m - T_0)\exp\left(-V\frac{c\rho}{\kappa}(x - Vt)\right), \qquad x > Vt, \quad t > 0, \tag{1.4}$$

with

$$V = \frac{H}{\rho[L + c(T_m - T_0)]} \tag{1.5}$$

(here c, ρ, κ, L, T_m, and H are positive constants).

The above expressions show that the speed of the wave is determined by the limit of the temperature for $x \to +\infty$, here T_0. Moreover, the speed does not depend on the thermal conductivity κ, which only influences the form of the temperature profile near the melting front.

Of course, any space or time translation ($x \to x + x_0$, $t \to t + t_0$) gives a new wave solution of our equations.

A rigorous mathematical analysis of the problem in a slab of finite thickness can be found in [6]. Here the authors prove the existence and the uniqueness of the solution. Conditions are also given under which the free boundary switches a finite number of time between $\dot{s} = 0$ and $\dot{s} < 0$. More results on the regularity of the boundary can be found in [7].

In this paper we are concerned with the problem of the stability of the traveling wave solutions (1.4) in the semi-infinite solid. We give a condition on the initial datum under which the solution is asymptotic to a traveling wave. This condition applies, in particular, to the case $T_0(x) \equiv \text{const}$. In the paper by Landau [8] the convergence to the traveling wave is assumed and the author is able to compute the phase shift of the limiting wave by an integral relation. Here we prove that the free boundary does converge to a traveling wave front, and we give an estimate of the difference. The approach to the problem is completely different from that of Landau. The first step is to transform the ablation problem into a Stefan problem (i.e., with Stefan free boundary condition) for a heat equation with convection. This is made by means of the transformation

$$\xi = x - vt, \quad u(\xi, t) = -(T(x,t) - T_m), \quad \sigma(t) = s(t) - vt, \tag{1.6}$$

with $v = H/\rho L$, which transforms the ablation problem into

$$\frac{\partial u}{\partial t} - v\frac{\partial u}{\partial \xi} - K\frac{\partial^2 u}{\partial \xi^2} = 0, \qquad x > \sigma(t), \quad t > 0, \tag{1.7}$$

$$u(\sigma(t), t) = 0, \qquad t > 0, \tag{1.8}$$

$$K\frac{\partial u}{\partial \xi}(\sigma(t), t) = -\lambda\dot{\sigma}(t), \qquad t > 0. \tag{1.9}$$

Now u is positive and we can consider problem (1.7)-(1.8) as a Stefan problem for the temperature of a liquid phase in the presence of a convection term. This problem admits traveling wave solutions for any $v \neq 0$, given by $\sigma(t) = -Wt$, and

$$u(\xi, t) = \frac{\lambda W}{v - W}\left\{\exp\left(-\frac{v - W}{K}(\xi + Wt)\right) - 1\right\}, \tag{1.10}$$

where W and $u_0 = \lim_{\xi \to +\infty} u(\xi)$ are related by $W = vu_0/(\lambda + u_0)$.

This solution is mathematically correct for any v but is bounded at $\xi = +\infty$ only for $v > 0$. This is in fact the case for the transformed problem where $v = H/\rho L$. The wave now moves toward the left but with a speed less (in modulus) than the convection flow v. Using (1.6) to return to the original variables, we recover the melting wave (1.4).

Problem (1.7)-(1.9) was also proposed by Paveri-Fontana as a limiting case for a model of the enclosed atomizer in analytical chemistry [9]. For this model he also considered the case of negative value of v. In this case no bounded wave solution exists, and the candidate asymptotic state is $u \equiv 0$ with a fixed free boundary, $s(t) \equiv b$.

2. Statement of the problem and results. Here we take $x = -\xi$ so that the equations read

$$u_t + vu_x - \kappa u_{xx} = 0, \qquad x < s(t), \quad t > 0, \tag{2.1}$$

$$s(0) = 0, \tag{2.2}$$

$$u(x, t) = u_0(x), \qquad x < 0, \tag{2.3}$$

$$u(s(t), t) = 0, \qquad t > 0, \tag{2.4}$$

$$\kappa u_x(s(t), t) = -\lambda \dot{s}(t), \qquad t > 0, \tag{2.5}$$

where $v \in \mathbb{R}$, $\kappa, \lambda > 0$, and $u_0(x) \geqq 0$.

Unless $u_0(x) \equiv 0$, we have $u(x, t) > 0$ and the Vyborny-Friedman boundary point principle [5], [10] ensures that $u_x(s(t), t) < 0$ for any $t > 0$. Consequently,

$$\dot{s}(t) > 0, \qquad t > 0. \tag{2.6}$$

The above inequality holds whatever is the sign of v.

When $v > 0$, taking

$$u_0(x) = u_\infty \left\{ 1 - \exp\left(\frac{v\lambda}{\kappa(\lambda + u_\infty)} x \right) \right\}, \qquad x < 0,$$

the solution of (2.1)-(2.5) is the traveling wave

$$u_W(x) = u_\infty \left\{ 1 - \exp\left(\frac{v\lambda}{\kappa(\lambda + u_\infty)} (x - Wt) \right) \right\}, \qquad x < Wt, \quad t > 0, \tag{2.7}$$

with speed

$$W = \frac{vu_\infty}{\lambda + u_\infty} \tag{2.8}$$

and free boundary $x = Wt$.

Note here that the speed of the free boundary is less than the speed of the convection drift and does not depend on the thermal diffusivity κ. The factor $u_\infty/(\lambda + u_\infty)$ depends on the value of $\lim_{x \to -\infty} u_0(x)$, and on the latent heat. If equations (2.1)-(2.5) originate from an ablation problem, then the corresponding solution in the original frame of reference is a backward traveling wave with speed $\lambda v/(\lambda + u_\infty)$.

The problem has a natural invariance with respect to space and time translations. In particular, if we change (2.2) into $s(0) = b$ and x into $x + b$, then the function $u_{W,b}(x, t) = u_W(x - b, t)$ is a traveling wave solution as well.

For $v = 0$, (2.1)-(2.5) admit a similarity solution corresponding to the initial datum $u_0 \equiv 1$ [10]. No special bounded solution is known by us for $v < 0$, except for the trivial one $u \equiv 0$, $s \equiv$ const.

The main result of this paper is the following asymptotic estimate for the free boundary.

THEOREM. *Let $v>0$ and suppose that $u_0(x)$ is a bounded function satisfying*

(i) $\lim_{x\to-\infty} u_0(x)=u_\infty$, $0<u_\infty<+\infty$,

(ii) $x[u_\infty-u_0(x)]\in L^1(-\infty,0)$.

Then there exists a constant C such that

$$|s(t)-Wt-b|<\frac{C}{1+t^{1/4}},\qquad W=\frac{vu_\infty}{\lambda+u_\infty}. \tag{2.9}$$

The phase shift b is given by

$$b=\frac{\kappa u_\infty}{\lambda v}+\frac{1}{\lambda+u_\infty}\int_{-\infty}^{0}(u_0(x)-u_\infty)\,dx. \tag{2.10}$$

Moreover, if the difference $u_\infty-u(x)$ decays exponentially at $-\infty$, then the difference $s(t)-Wt-b$ decays exponentially in time as well.

A similar result holds for v negative. In this case the asymptotic state is $u\equiv 0$ and the free boundary has a finite limit as $t\to\infty$.

3. Proof of the theorem. We make use of a smoothing procedure which transforms the original problem into a quasi-variational problem in a fixed domain (see [1], [11]).

The transformed problem is referred to as the oxygen-consumption problem [3].

We look for a function $s(t)$ defined for $t>0$ and for a function $c(x,t)$ defined in $x<s(t)$, $t>0$, such that

$$c_t+vc_x-\kappa c_{xx}+\lambda=0,\qquad x<s(t),\quad t>0, \tag{3.1}$$

$$c(x,0)=c_0(x),\qquad x<0=s(0), \tag{3.2}$$

$$c(s(t),t)=c_x(s(t),t)=0,\qquad t>0, \tag{3.3}$$

where the initial datum c_0 is the solution of the ordinary differential equation

$$c''-\frac{v}{\kappa}c'-\frac{\lambda}{\kappa}=\frac{u_0(x)}{\kappa},\qquad x<0, \tag{3.4}$$

$$c_0(0)=c_0'(0)=0,$$

and u_0 is the initial datum for u in (2.3).

The function $u(x,t)=c_t(x,t)$, together with the free boundary $s(t)$, gives the solution of the Stefan problem (2.1)-(2.5).

The problem may be restated eliminating any direct reference to the free boundary. If we extend the solution of (3.1)-(3.3) on the right of the free boundary to be identically zero, the function $c(x,t)$ so defined in $\mathbb{R}\times\mathbb{R}^+$ is the unique nonnegative solution of

$$c_t+vc_x-\kappa c_{xx}+\lambda H(c)=0\quad\text{in }\mathbb{R}\times\mathbb{R}^+, \tag{3.5}$$

$$c(x,t)=c_0(x)\quad\text{in }\mathbb{R}, \tag{3.6}$$

where c_0 is the solution of (3.4) for $x<0$ and it is extended to vanish identically for $x>0$, and

$$H(\eta)=\begin{cases}1, & \eta>0,\\ 0, & \eta\leqq 0.\end{cases} \tag{3.7}$$

Here the free boundary $x = s(t)$ is recovered a posteriori from $c(x, t)$ as $s(t) = \sup\{x \in \mathbb{R}: c(x, t) > 0\}$. On $x = s(t)$ we have $c = c_x = c_t = 0$ but

$$\lim_{x \uparrow s(t)} c_{xx} = \frac{\lambda}{\kappa}. \tag{3.8}$$

We will use the function c to establish some results on the asymptotic behavior of $s(t)$. The major advantage of this formulation of the problem is the possibility of comparing the solutions of (3.5) corresponding to different initial values.

The solution of (3.4) for a given $u_0(x)$ is given by

$$c(x, 0) = \left[-\frac{\lambda + u_\infty}{v} \left(x + \frac{\kappa}{v}(1 - e^{(v/\kappa)x}) \right) + \frac{1}{\kappa} \int_0^x \left[e^{(v/\kappa)y} \int_0^y e^{-(v/\kappa)\xi} g(\xi)\, d\xi \right] dy \right]_+, \tag{3.9}$$

where $g(x) = u_0(x) - u_\infty$, and $[s]_+ = \max\{0, s\}$. Note that $c(x, 0) \equiv 0$ for $x > 0$.

In the case of a traveling wave $u_{W,b}(x, t)$, the transformed c is a traveling wave of (3.5)-(3.6), namely,

$$c_{W,b}(\xi) = -\frac{\lambda + u_\infty}{v} \left[\xi + \frac{\lambda + u_\infty}{v} \frac{\kappa}{\lambda} \left(1 - \exp\left(\frac{v}{\kappa} \frac{\lambda}{\lambda + u_\infty} \xi \right) \right) \right]_+, \tag{3.10}$$

with $\xi = x - Wt - b$, and W given by (2.8).

We are now in position to prove the theorem.

LEMMA. *Suppose that $u_0(x)$ satisfies assumptions* (i) *and* (ii) *of the theorem. Then by taking b as in* (2.10), *the difference $c(x, 0) - c_{W,b}(x, 0)$ belongs to $L^1(\mathbb{R})$.*

Proof. We have $c - c_{W,b} \equiv 0$ for $x > \max\{0, b\}$ and

$$c(x, 0) - c_{W,b}(x, 0) = -c_{W,b}(x, 0), \qquad 0 < x < b, \quad \text{if } b > 0, \quad \text{or}$$

$$c(x, 0) - c_{W,b}(x, 0) = c(x, 0), \qquad b < x < 0, \quad \text{if } b < 0, \quad \text{and}$$

$$c(x, 0) - c_{W,b}(x, 0) = \frac{\lambda + u_\infty}{v} \left[\frac{\kappa}{v}(e^{(v/\kappa)x} - 1) - b + \frac{\lambda + u_\infty}{v} \frac{\kappa}{\lambda} \cdot \left(1 - \exp\left(\frac{v}{\kappa} \frac{\lambda}{\lambda + u_\infty} (x - b) \right) \right) \right]$$

$$+ r(x), \quad \text{for } x < \min\{0, b\},$$

where

$$r(x) = \int_0^x \left[e^{(v/\kappa)y} \int_0^y e^{-(v/\kappa)\xi} g(\xi)\, d\xi \right] dy, \tag{3.11}$$

and

$$r_\infty = \lim_{x \to -\infty} r(x). \tag{3.12}$$

First note that r_∞ is finite because hypotheses (i) and (ii) of the theorem imply that $g \in L^1(\mathbb{R}^-)$. In fact, exchanging the order of integration in (3.11), we have

$$r(x) = \frac{\kappa}{v} \int_0^x e^{(v/\kappa)(x-\xi)} g(\xi)\, d\xi - \frac{\kappa}{v} \int_0^x g(\xi)\, d\xi$$

(note that $x < 0$ and $\xi > x$ in the first integral) and then

$$r_\infty = \frac{\kappa}{v} \int_{-\infty}^0 g(\xi)\, d\xi.$$

Now to prove that $c - c_{W,b} \in L^1(\mathbb{R})$, we must verify that $r(x) - r_\infty$ belongs to $L^1(\mathbb{R}^-)$. First we rewrite $r(x) - r_\infty$ as

$$
\begin{aligned}
r(x) - r_\infty = & \frac{\kappa}{v} \int_x^0 e^{(v/\kappa)(x-\xi)} g(\xi)\, d\xi \\
& - \frac{\kappa}{v} \int_{-\infty}^x g(\xi)\, d\xi + \lim_{Q \to \infty} \frac{\kappa}{v} \int_{-Q}^x e^{-(v/\kappa)(Q+\xi)} g(\xi)\, d\xi.
\end{aligned} \tag{3.13}
$$

The first term in (3.13) is integrable since r_∞ is finite (see (3.11)). The third term is dominated by the second one (if g is positive, if not, simply consider $|g|$). So it remains to prove that

$$
\int_{-\infty}^x g(\xi)\, d\xi \in L^1(\mathbb{R}^-).
$$

This is true if $x \cdot g(x)$ is integrable in $(-\infty, 0)$, i.e., under hypothesis (ii). □

With our assumptions on the initial datum $u_0(x)$, and with the phase b given by (2.10), we have, for $x \to -\infty$

$$
|\delta(x, 0)| = |c(x, 0) - c_{w,b}(x, 0)| \leqq \text{const.}\, (e^{Cx} + |r_\infty - r(x)|), \tag{3.14}
$$

so the difference of the two solutions for $t = 0$ belongs to $L^1(\mathbb{R})$. Moreover, the difference satisfies the linear nonhomogeneous parabolic equation

$$
\delta_t + v\delta_x - \kappa\delta_{xx} = \lambda(H(c_{w,b}) - H(c)) = f(x, t). \tag{3.15}
$$

The nonhomogeneous term in (3.15) is different from zero only where one and only one of the two solutions of (3.5) vanishes, i.e., in

$$
N_\delta = \{(x, t)\colon 0 = c(x, t) \cdot c_{w,b}(x, t) < |c(x, t) - c_{w,b}(x, t)|\}. \tag{3.16}
$$

Moreover, in N_δ we always have

$$
\delta(x, t) \cdot f(x, t) < 0, \tag{3.17}
$$

which implies that

$$
|\delta(x, t)| \leqq \varepsilon(x, t), \tag{3.18}
$$

where ε is the solution of

$$
\varepsilon_t + v\varepsilon_x - \kappa\varepsilon_{xx} = 0, \qquad \varepsilon(x, 0) = |\delta(x, 0)|. \tag{3.19}
$$

The function $\varepsilon(x, t)$ is given by

$$
\varepsilon(x, t) = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{+\infty} e^{-(x+vt-\xi)^2/4\kappa t} |\delta(\xi, 0)|\, d\xi \tag{3.20}
$$

from which we get an estimate for δ,

$$
|\delta(x, t)| \leqq \text{const.} \frac{1}{1 + \sqrt{t}} \quad \text{uniformly in } x \in \mathbb{R}. \tag{3.21}
$$

Finally, from (3.21) we can deduce an estimate for the difference of the two free boundaries. In fact, from the jump relation (3.8) for the second derivative c_{xx} on the free boundary it follows that both $c_{W,b}(x, t)$ and $c(x, t)$ behave like parabolae in the vicinity of their free boundaries.

Now suppose for instance that $s(t) < Wt + b$ for a given large t. Then at $x = s(t)$ we have $\delta(s(t), t) = c_{W,b}(s(t), t) < \text{const.}\, 1/(1+\sqrt{t})$. Since the solutions c's are monotone decreasing functions and $c_{W,b}(x, t) = (\lambda/\kappa)(\xi + b)^2 + O(\xi + b)^3 > (\lambda/2\kappa)(\xi + b)^2$ for $\bar{\xi} < \xi < 0$, $\xi = x - Wt$, then for t sufficiently large we have $(\lambda/2\kappa)(s(t) - Wt + b)^2 < c_{W,b}(s(t), t) < \text{const.}\, 1/(1+\sqrt{t})$, so that a constant exists such that

$$|s(t) - Wt - b| \leqq \text{const.} \frac{1}{1 + t^{1/4}}. \tag{3.22}$$

The same conclusion holds if $Wt + b < s(t)$.

A far better estimate holds if the difference $\delta(x, 0)$ decays exponentially as $x \to -\infty$, i.e., if the difference $r_\infty - r(x)$ behaves as does $e^{\gamma x}$ for some positive γ. So now we assume that

$$|\delta(x, 0)| \leqq \text{const.}\, e^{ax} \quad \text{with } a = \min\left\{\frac{v\lambda}{\kappa(\lambda + u_\infty)}, \gamma\right\}. \tag{3.23}$$

We consider the problem in the moving frame of reference with speed W, where the traveling wave is at rest (and the free boundary of our solution has a finite limit). By the same argument used above we control the difference δ by a solution $\varepsilon(y, t)$ of

$$\varepsilon_t + (v - W)\varepsilon_y - \kappa\varepsilon_{yy} = 0, \quad y = x - Wt, \quad \varepsilon(y, 0) \geqq \text{const.}\, e^{ax}. \tag{3.24}$$

For $\varepsilon(y, t)$ we can now choose the explicit wave type solution of (3.24), namely,

$$\varepsilon(y, t) = \text{const.}\, e^{\alpha(y - \beta t)}, \tag{3.25}$$

where α and β can be chosen to satisfy

$$0 < \alpha < a \quad \text{and} \quad \beta = v - W - \kappa\alpha > 0. \tag{3.26}$$

Then the difference δ decays exponentially in any compact subset, in particular, in a fixed neighborhood of the limit of the free boundary (the phase b). Recalling again (3.8), we get

$$|s(t) - Wt - b| \leqq \text{const.}\, e^{-\alpha\beta t/2}. \tag{3.27}$$

Remark. A similar analysis can be done in the case where v is negative. Now the asymptotic state for suitable initial data is a trivial solution of the problem with $u \equiv 0$ and $s(t) \equiv \text{const}$. Again we can prove convergence in the L^∞ norm of the function c's solutions of problem (3.1)-(3.4), and then we can deduce the estimate for the convergence of the free boundaries.

Acknowledgment. The author is greatly indebted to Stefano L. Paveri-Fontana for suggesting this problem and for various discussions.

REFERENCES

[1] C. BAIOCCHI, *Variational inequalities and free boundary problems*, in Variational Inequalities and Complementarity Problems, R. W. Cottle, F. Giannessi, and J. L. Lions, eds., John Wiley, New York, 1980.

[2] B. A. BOLEY, *An applied overview of moving boundary problems*, in Moving Boundary Problems, D. G. Wilson, A. D. Solomon, and P. T. Boggs, eds., Academic Press, New York, 1978.

[3] J. CRANK AND R. S. GUPTA, *A moving boundary problem arising from the diffusion of oxygen in absorbing tissue*, J. Inst. Math. Appl., 9 (1972), pp. 19-33.

[4] C. M. ELLIOTT AND J. R. OCKENDON, *Weak and Variational Methods for Moving Boundary Problems*, Res. Notes in Math., 59, Pitman, London, 1982.

[5] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[6] A. FRIEDMAN AND L. S. JIANG, *A Stefan-Signorini problem*, J. Differential Equations, 51 (1984), pp. 213-131.

[7] L. S. JIANG, *Remarks on the Stefan-Signorini problem*, in Free Boundary Problems: Applications and Theory Vol. III, A. Bossavit, A. Damfamian, and M. Fremond, eds., Res. Notes in Math., 120, Pitman, London, 1985, pp. 13-19.

[8] H. G. LANDAU, *Heat conduction in a melting solid*, Quart. J. Appl. Math., 8 (1950), pp. 81-94.

[9] S. PAVERI-FONTANA, *Traveling waves for the enclosed atomizer*, Math. Modelling, 8 (1987), pp. 275-278.

[10] L. I. RUBINSTEIN, *The Stefan Problem*, Trans. Math. Monthly American Mathematical Society, Providence, RI, 1971.

[11] A. SCHATZ, *Free boundary problems of Stefan type with prescribed flux*, J. Math. Anal. Appl., 28 (1969), pp. 569-580.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1394-1414, November 1990
© 1990 Society for Industrial and Applied Mathematics
003

ASYMPTOTIC ANALYSIS FOR A STIFF VARIATIONAL PROBLEM ARISING IN MECHANICS*

GABRIEL NGUETSENG†

Abstract. A new approach in the theory of homogenization is carried out on the variational boundary value problem of the stiff type that governs the small vibrations of a periodic mixture of an elastic solid and a slightly viscous fluid. A convergence theorem is proved, which gives the behaviour of the solution and points out the role of the connectedness of phases in the mechanics of mixtures.

Key words. homogenization, convergence, periodic, mixtures of fluids and solids

AMS(MOS) subject classifications. 35B40, 35B25

1. Preliminaries.

1.1. Introduction. The recent developments of the theory of homogenization have considerably formed the knowledge of the periodic heterogeneous media. Besides the works [2] and [18] in which the methods are initiated and a great variety of problems are studied, many papers have been written in recent years on various aspects of homogenization because of their outstanding importance in physical applications. Without pretensions of an exhaustive bibliography, we refer the reader to Cioranescu and Murat [3], Cioranescu and Saint-Jean Paulin [4], Tartar [23], for perforated domains; Tartar [22], Ene and Sanchez-Palencia [6], Conca [5], for specific problems in fluid mechanics.

There is another field of particular interest in physical applications, namely, that dealing with mixtures in mechanics. Studies can be found, for example, in Levy [13], Fleury [7], and Nguetseng [15] (see also Sanchez-Palencia [18, Chap. 8]). From the mathematical point of view this field has not yet been extensively studied. In many situations the validation of the formal analysis, that is, the convergence of the homogenization process, remains an open question.

The study of periodic mixtures of elastic solids and compressible viscous fluids reveals that very different behaviours occur according to whether or not the total fluid region is connected, and according to the orders of magnitude of the elasticity and viscosity coefficients. From Sanchez-Hubert [20] we know that, whether or not the total fluid part is connected, if the elasticity coefficients together with the viscosity coefficients are $O(1)$ (with respect to ε, the period of the structure), then the limit of the (perturbed) displacement vector does not depend on the local variables. Here, the convergence of the homogenization process is proved after the energy method [2].

The very complex situation, which we study here, is that in which the elasticity coefficients are $O(1)$, whereas the viscosity coefficients are $\varepsilon^2 O(1)$. Under these assumptions, the formal analysis in [13] (see also [17] and [18, Chap. 8]) leads to the following proposition. If the fluid cell (that part of the fluid lying in the period of reference) is strictly contained in the period of reference (see Fig. 1)—which implies that the total fluid part is not connected—the formal limit of the displacement in the mixture does not depend on the local variables. If, however, it "touches" each face of the basic period and moreover the corresponding total fluid part is connected (the model represented in Fig. 2 falls under this framework), the formal limit of the displacement depends on the local variables. Thus the problem is a member of a specific family of

* Received by the editors August 2, 1988; accepted for publication (in revised form) December 1, 1989.
† Department of Mathematics, University of Yaounde, P.O. Box 812,Yaounde, Cameroon.

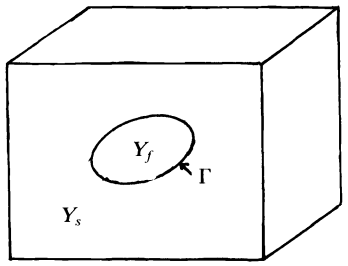


FIG. 1

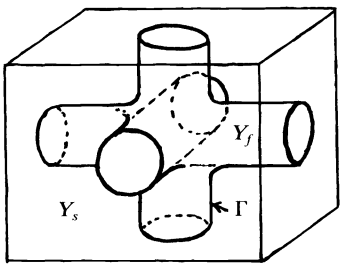


FIG. 2

unusual homogenization problems referred to as singular homogenization problems [16]. In both cases the convergence of the homogenization process remains open. As was mentioned in [16], the energy method is not flexible enough to handle the present situation.

Our main result is the proof of convergence in the above situation. More precisely, for each model considered above, we rigorously prove the convergence of "the problem in ε" to the homogenized problem, and validate the preceding formal proposition.

1.2. The mathematical problem. We start with some basic notation. Let the three-dimensional Euclidean space $\mathbb{R}^3$ of the variables $y=(y_1, y_2, y_3)$ be regarded as a periodic set with period of reference $Y=]-\frac{1}{2}, \frac{1}{2}[^3$. In what follows, Y is decomposed as

$$Y = Y_s \cup Y_f \cup \Gamma, \tag{1.1}$$

where Y_s and Y_f are connected open sets in $\mathbb{R}^3$ and Γ is the smooth surface that separates them. We will denote by $\tilde{Y}_s$ (respectively, $\tilde{Y}_f$) the Y-periodic continuation of Y_s (respectively, Y_f), that is, the union of all the $(Y_s \cup (\bar{Y}_s \cap \partial Y)) + k$, k ranging over Z^3 (and an analogous formula for $\tilde{Y}_f$) where $\bar{Y}_s$ is the closure of Y_s in $\mathbb{R}^3$ and ∂Y the boundary of the cube Y.

In the sequel two models are considered:

(i) Y_f is a smooth open set with its closure contained in Y (Fig. 1).

(ii) Y_f is made of three tubes with same radius R (with R sufficiently small compared to $\frac{1}{2}$) and same length 1, whose axes coincide with the coordinate axes, respectively. Moreover, the intersection of the tubes is smoothed down so that the interface Γ is sufficiently smooth (Fig. 2). Note that the origin of the space $\mathbb{R}^3$ is the symmetry center of Y_f.

In both cases the Y-periodic continuation $\tilde{Y}_s$ (respectively, $\tilde{Y}_f$) of Y_s (respectively, Y_f) is an open subset of $\mathbb{R}^3$ with smooth boundary $\tilde{\Gamma}$ (the Y-periodic continuation of Γ). In the case (i) $\tilde{Y}_f$ is not connected, whereas $\tilde{Y}_s$ is. In the case (ii) both $\tilde{Y}_f$ and $\tilde{Y}_s$ are connected.

Now in the space $\mathbb{R}^3$ of the variables $x=(x_1, x_2, x_3)$ we consider a smooth bounded open set Ω (with boundary $\partial\Omega$) and we set

$$\Omega_\varepsilon^s = \Omega \cap \varepsilon \tilde{Y}_s, \qquad \Omega_\varepsilon^f = \Omega \cap \varepsilon \tilde{Y}_f, \tag{1.2}$$

where ε $(0<\varepsilon<1)$ is a sequence tending to zero.

Furthermore, we introduce the following notation based on vector functions such as $v=(v^i(x))$ and $w=(w^i(y))$:

$$E_{ij}(v) = \frac{1}{2}\left(\frac{\partial v^i}{\partial x_j} + \frac{\partial v^j}{\partial x_i}\right), \tag{1.3}$$

$$e_{ij}(w)=\frac{1}{2}\left(\frac{\partial w^i}{\partial y_j}+\frac{\partial w^j}{\partial y_i}\right). \tag{1.4}$$

Correspondingly, div_y denotes the divergence operator with respect to y, while div_x (or simply div, when there is no danger of confusion) denotes the same operator with respect to x.

Finally, if V is a vector space (because of the use of the Laplace transform in the sequel, all the vector spaces are considered over the complex field $\mathbb{C}$), the vector space of the same name written in boldface represents the corresponding product space $V^3 = V \times V \times V$.

We are now in a position to state the mathematical problem. The summation convention is used throughout the rest of this section.

Let a_{ijkh} $(1 \leqq i, j, k, h \leqq 3)$ and η, μ be real numbers subject to the following conditions:

$$a_{ijkh} = a_{jikh} = a_{ijhk} = a_{khij}, \tag{1.5}$$

$$a_{ijkh}\,\xi_{kh}\,\xi_{ij} \geqq c\xi_{ij}\xi_{ij} \quad (c>0) \quad \forall \xi_{ij} \in \mathbb{R}, \quad \xi_{ij} = \xi_{ji}, \tag{1.6}$$

$$\mu > 0, \quad \frac{\eta}{\mu} \geqq -\frac{2}{3}\alpha \quad \text{with } 0<\alpha<1. \tag{1.7}$$

Next, we define the Y-periodic functions (i.e., periodic with period 1 in each coordinate):

$$c_{ijkh}(y) = a_{ijkh} \quad \text{in } Y_s, \qquad \gamma\delta_{ij}\delta_{kh} \quad \text{in } Y_f,$$

$$b_{ijkh}(y) = 0 \quad \text{in } Y_s, \qquad \eta\delta_{ij}\delta_{kh} + \mu(\delta_{ik}\delta_{jh} + \delta_{ih}\delta_{jk}) \quad \text{in } Y_f,$$

(δ_{ij} is the Kronecker symbol)

$$\rho(y) = \rho^s \quad \text{in } Y_s, \qquad \rho^f \quad \text{in } Y_f,$$

where ρ^s, ρ^f are positive constants, and $\gamma = c_0^2\rho^f$, $c_0>0$ (the physical definition of those constants is given in the sequel).

Now we introduce the sequence ε $(0<\varepsilon<1)$ and, with the standard notation

$$w^\varepsilon(x) = w\left(\frac{x}{\varepsilon}\right) \quad \text{for } w = w(y), \tag{1.8}$$

we define, for $u=(u^i)$ and $v=(v^i)$ in $\mathbf{H}^1(\Omega)$, the forms:

$$b^\varepsilon(u, v) = \int_\Omega b^\varepsilon_{ijkh}\frac{\partial u^k}{\partial x_h}\overline{\frac{\partial v^i}{\partial x_j}}\,dx,$$

$$c^\varepsilon(u, v) = \int_\Omega c^\varepsilon_{ijkh}\frac{\partial u^k}{\partial x_h}\overline{\frac{\partial v^i}{\partial x_j}}\,dx.$$

Finally, for a given $f=(f^i)$ in $L^2_{\mathrm{loc}}(0, +\infty; \mathbf{L}^2(\Omega))$, f independent of ε and satisfying

$$\|f(t)\|^2_{\mathbf{L}^2(\Omega)} \leqq K e^{mt} (K>0, m \in \mathbb{R}) \quad \text{for almost all } 0<t<+\infty, \tag{1.9}$$

we consider, for fixed ε, the variational initial value problem (see [18, Chap. 8]):

(1.10) Find u_ε, function of t with values in $\mathbf{H}_0^1(\Omega)$ such that

$$\int_\Omega \rho^\varepsilon \frac{\partial^2 u_\varepsilon^i}{\partial t^2}\overline{v^i}\,dx+\varepsilon^2 b^\varepsilon\left(\frac{\partial u_\varepsilon}{\partial t}, v\right)+c^\varepsilon(u_\varepsilon, v)=\int_\Omega f^i\overline{v^i}\,dx$$

for all $v=(v^i)$ in $\mathbf{H}_0^1(\Omega)$,

$$u_\varepsilon(0)=\frac{\partial u_\varepsilon}{\partial t}(0)=0.$$

Remark 1. The sesquilinear forms b^ε and c^ε can be written explicitly as follows:

$$b^\varepsilon(u, v)=\int_{\Omega_\varepsilon^f}(\eta \operatorname{div} u\,\overline{\operatorname{div} v}+2\mu E_{ij}(u)\overline{E_{ij}(v)})\,dx,$$

$$c^\varepsilon(u, v)=\int_{\Omega_\varepsilon^s} a_{ijkh}\frac{\partial u^k}{\partial x_h}\frac{\overline{\partial v^i}}{\partial x_j}\,dx+\int_{\Omega_\varepsilon^f}\gamma \operatorname{div} u\,\overline{\operatorname{div} v}\,dx,$$

where we recall that, by virtue of (1.5), the following holds:

$$\int_{\Omega_\varepsilon^s} a_{ijkh}\frac{\partial u^k}{\partial x_h}\frac{\partial v^i}{\partial x_j}\,dx=\int_{\Omega_\varepsilon^s} a_{ijkh}E_{kh}(u)E_{ij}(v)\,dx. \tag{1.11}$$

For fixed ε, problem (1.10) is the variational formulation of the initial boundary value problem that governs the small vibrations (in the framework of the linearized theory of small perturbations) of a solid-fluid mixture, the geometric structure of which is εY-periodic. The constants a_{ijkh} are the elasticity coefficients of the solid, which is a homogeneous body (but this property is not essential), with the classical symmetry and positivity conditions, (1.5) and (1.6), respectively. The viscosity coefficients of the fluid are $\mu\varepsilon^2$ and $\eta\varepsilon^2$. They satisfy (1.7), whose physical justification is to be found in [12]. The positive numbers ρ^s, ρ^f are the densities of the mass of the solid and the fluid, respectively, in the reference state at rest; and c_0 is the velocity of sound.

In the perturbed state, the partial differential equations governing the motion of the mixture are written in the bounded open set Ω for the displacement vector u_ε, with respect to the reference state (the elasticity equations are written in Ω_ε^s, while the fluid dynamic equations are written in Ω_ε^f), with homogeneous Dirichlet boundary condition and initial data, and homogeneous transmission conditions for the continuity of displacement and stress at the interface between Ω_ε^f and Ω_ε^s. The vector function f is a given force, independent of ε. See [13] and [18] for further details.

Conditions (1.6) and (1.7) ensure the existence and uniqueness in problem (1.10) for fixed ε (see, e.g., [18, Chap. 8]). The proof utilizes the Laplace transform (see [8], [9], [21]), the function $f=(f^i)$, and the unknown $u_\varepsilon=(u_\varepsilon^i)$ being considered as defined for $-\infty<t<+\infty$ with supports in $[0, +\infty[$. Note that although none of the sesquilinear forms b^ε and c^ε is coercive (on $\mathbf{H}_0^1(\Omega)$), the form $\lambda b^\varepsilon+c^\varepsilon$ is coercive on $\mathbf{H}_0^1(\Omega)$ for all $\lambda\in\mathbb{C}$, $\operatorname{Re}\lambda>0$.

Existence and uniqueness in (1.10) can also be obtained after the fashion of [20] by semigroups.

The corresponding homogenization problem is analysed formally in [13], [18, Chap. 8] by means of the classical method using multiple-scale asymptotic expansions. It remains to prove the convergence of the preceding homogenization process. This we accomplish by an adaptation of the idea in [16] and use of an appropriate "next order approximation theorem" analogous to Theorem 3 in the above paper. The Laplace transform will prove very useful in the sequel. Denoting by $\hat{v}$ (function of the variable

λ) the Laplace transform of a function $v(t)$, problem (1.10) becomes for fixed λ, $\operatorname{Re}\lambda > \lambda_0 > 0$ (λ_0 large enough):

(1.12) Find $\hat{u}_\varepsilon \in \mathbf{H}_0^1(\Omega)$ such that

$$\lambda^2 \int_\Omega \rho^\varepsilon \hat{u}_\varepsilon^i \overline{v^i}\, dx + \lambda\varepsilon^2 b^\varepsilon(\hat{u}_\varepsilon, v) + c^\varepsilon(\hat{u}_\varepsilon, v) = \int_\Omega \hat{f}^i \overline{v^i}\, dx$$

for all $v = (v^i)$ in $\mathbf{H}_0^1(\Omega)$,

where the argument, λ, in $\hat{u}_\varepsilon(\lambda)$ and $\hat{f}(\lambda)$, is omitted.

Our further analysis will be based on (1.12) (the "stationary version") which is more convenient. Under this form it is apparent that problem (1.10), as $\varepsilon \downarrow 0$, is a perturbation problem of the stiff type [14].

This paper is organized as follows. In § 2 we prove a convergence theorem in view of the study of homogenization problems (in the framework of periodic structures) on variable domains (e.g., Ω_ε^s). Although this theorem is invoked in a particular context, it can be used in some other fields in homogenization (for periodic structures) dealing with variable domains (perforated materials, fluid flow in porous media $\cdots$; see the references in § 1.1). To this end, the theorem is presented in as general a form as possible.

Section 3 deals with the actual analysis of the problem, both of the models (i) and (ii) (relative to Fig. 1 and Fig. 2, respectively) being considered. They are studied simultaneously (our approach offers this possibility). However, the results specific to each case are pointed out in the sequel.

In § 4 we derive the homogenized problem and show the uniqueness of its solution. Next we prove the convergence, as $\varepsilon \downarrow 0$, of u_ε (the solution of (1.10)) to the preceding solution.

Finally, the Appendix is devoted to an extension result. Indeed, in order to have good estimates in certain homogenization problems dealing with variable domains it is often essential to suitably extend any function, say, of class H^1 on Y_s, to a function of the same class on the whole period of reference Y. For the model represented in Fig. 1, a suitable extension operator is available (see, e.g., [4], [5]). The cases in which the inclusion (e.g., Y_f) is not strictly contained in Y (and, in particular, the case of Fig. 2) are, in general, unresolved.

2. A general convergence result.

2.1. Notation and preliminaries. In the Euclidean space $\mathbb{R}^N$ ($N \geqq 2$) of the variables $y = (y_1, \cdots, y_N)$ we consider an open set Y_0 contained in the cube $Y =]-\frac{1}{2}, +\frac{1}{2}[^N$ and subject to the following conditions:

(2.1) For any j ($1 \leqq j \leqq N$) the sets $\partial Y_0 \cap \{y;\ y_j = -\frac{1}{2}\}$ and $\partial Y_0 \cap \{y;\ y_j = \frac{1}{2}\}$ (∂Y_0 denotes the boundary of Y_0) are of nonzero measures (with respect to the $(N-1)$-dimensional measure). Moreover, they are symmetrical with respect to the plane $\{y;\ y_j = 0\}$;

(2.2) Y_0 contains a cylinder Q_j of length 1, whose axis is parallel to the y_j-axis ($j = 1, \cdots, N$);

(2.3) If $Y_0 \neq Y$ we denote by Γ the separation surface between Y_0 and the interior of $Y \setminus Y_0$, and we assume that Γ is smooth.

Remark 2. There is no condition concerning the intersection of the cylinders Q_j ($1 \leqq j \leqq N$).

Remark 3. The above conditions correspond to very practical situations. For example, $Y_0 = Y$, $Y_0 = Y \setminus K$ where K is the closure of a smooth bounded open set strictly contained in Y (e.g., Fig. 1), $Y_0 = Y_s$ (respectively, Y_f) in Fig. 2.

In the sequel $\tilde{Y}_0$ (respectively, $\tilde{Q}_j$) denotes the Y-periodic continuation of Y_0 (respectively, Q_j) (see § 1 for this definition). Note that $\tilde{Y}_0$ is a connected open set in $\mathbb{R}^N$ with smooth boundary $\tilde{\Gamma}$. On the other hand, $\tilde{Q}_j$ is a union of infinite cylinders contained in $\tilde{Y}_0$ (with same radius, and with axes parallel to the y_j-axis) and periodically distributed in the space.

It is essential to observe that the open sets $\tilde{Q}_j$ possess the useful property of "primitive preserving with respect to y_j" ($1 \leqq j \leqq N$); that is, e.g., for $j = 1$: a function w in $\mathcal{D}(\tilde{Q}_1)$ (the usual space of C^∞ functions with compact supports in $\tilde{Q}_1$) possesses a primitive with respect to y_1 in $\mathcal{D}(\tilde{Q}_1)$ if and only if it satisfies

$$\int_{-\infty}^{+\infty} w(t, y')\, dt = 0 \quad \text{for any } y' = (y_2, \cdots, y_N).$$

Some function spaces. We will denote by:

C_p^∞ the space of Y-periodic C^∞ functions on $\mathbb{R}^N$,

$\mathcal{D}_p(Y_0)$ $(Y_0 \neq Y)$ the space of functions $w \in C_p^\infty$, with support in $\tilde{Y}_0$ and for which there is some $r > 0$ (r depending on w) such that $w(y) = 0$ for all $y \in \tilde{Y}_0$, $d(y, \tilde{\Gamma}) \leqq r$ (where $\tilde{\Gamma}$ designates the boundary of $\tilde{Y}_0$ and d the Euclidean metric),

$L_p^2(Y_0)$ the space of $w \in L^2_{\text{loc}}(\tilde{Y}_0)$, w Y-periodic (which is a Hilbert space with the $L^2(Y_0)$-norm),

$H_p^1(Y_0)$ the space of $w \in L_p^2(Y_0)$, $\partial w / \partial y_i \in L_p^2(Y_0)$ for $i = 1, \cdots, N$ (which is a Hilbert space with the $H^1(Y_0)$-norm),

$H_p^1(Y_0)/\mathbb{C}$ the space of $w \in H_p^1(Y_0)$, $\int_{Y_0} w\, dy = 0$; on which the $H^1(Y_0)$-norm is equivalent to the norm

$$\|w\|_{H^1(Y_0)/\mathbb{C}} = \left(\sum_{i=1}^N \left\| \frac{\partial w}{\partial y_i} \right\|^2_{L^2(Y_0)} \right)^{1/2}.$$

In the case $Y_0 = Y$, we will write L_p^2 (respectively, H_p^1) in place of $L_p^2(Y)$ (respectively, $H_p^1(Y)$).

2.2. Statement of the theorem. In the space $\mathbb{R}^N$ of the variables $x = (x_1, \cdots, x_N)$ we consider a smooth bounded open set Ω. Next, we introduce ε $(0 < \varepsilon < 1)$, destined to tend to zero, and we define $\Omega_\varepsilon^0 = \Omega \cap \varepsilon Y_0$, where Y_0 satisfies (2.1)-(2.3). But in what follows we suppose $Y_0 \neq Y$ (the case $Y_0 = Y$ was studied in [16]; see Remark 5 below).

THEOREM 1. *Let $v_\varepsilon \in H^1(\Omega)$. Assume there exists a constant $c > 0$ (c independent of ε) such that*

$$\|v_\varepsilon\|_{L^2(\Omega)} \leqq c \quad \textit{for all } \varepsilon, \tag{2.4}$$

$$\sum_{i=1}^N \int_{\Omega_\varepsilon^0} \left| \frac{\partial v_\varepsilon}{\partial x_i} \right|^2 dx \leqq c \quad \textit{for all } \varepsilon. \tag{2.5}$$

Then we can extract a subsequence from ε (still denoted by ε for simplicity) such that, as $\varepsilon \downarrow 0$, we have

$$v_\varepsilon \to \tilde{v}_0 \quad \textit{in } L^2(\Omega)\textit{-weak}, \tag{2.6}$$

$$\int_\Omega v_\varepsilon w^\varepsilon \phi\, dx \to \int_{\Omega \times Y} v_0(x, y) w(y) \phi(x)\, dx\, dy, \tag{2.7}$$

$$\int_{\Omega_\varepsilon^0} \frac{\partial v_\varepsilon}{\partial x_i} w^\varepsilon \phi \, dx \to \int_{\Omega \times Y_0} \left(\frac{\partial u}{\partial x_i}(x) + \frac{\partial u_1}{\partial y_i}(x, y) \right) w(y) \phi(x) \, dx \, dy \tag{2.8}$$

for $i = 1, \cdots, N$, $\forall w \in L_p^2$, $\forall \phi \in \mathscr{K}(\bar{\Omega})$

(*the set of all restrictions to* Ω *of continuous functions on* $\mathbb{R}^N$ *with compact supports*), *where* $v_0 \in L^2(\Omega, L_p^2)$, $v_0(x, y) = u(x) + u_r(x, y)$ *with* $u \in H^1(\Omega)$, $u_r(x, y) = 0$ *almost everywhere in* Y_0 *for almost all* $x \in \Omega$, $u_1 \in L^2(\Omega; H_p^1(Y_0)/\mathbb{C})$, $\tilde{v}_0(x) = \int_Y v_0(x, y) \, dy$ (*mean value of* $v_0(x, \cdot)$).

Moreover, if $v_\varepsilon \in H_0^1(\Omega)$ (*i.e.,* $v_\varepsilon = 0$ *on* $\partial\Omega$) *then* $u \in H_0^1(\Omega)$.

We should recall the fundamental convergence result [16] on which the proof of Theorem 1 will be based.

LEMMA 1. *Let* $v_\varepsilon \in L^2(\Omega)$ (Ω *is any bounded open set in* $\mathbb{R}^N$) *such that*

$$\| v_\varepsilon \|_{L^2(\Omega)} \leqq c \quad \textit{for all } \varepsilon.$$

Then we can extract a subsequence from ε *such that, as* $\varepsilon \downarrow 0$

$$\int_\Omega v_\varepsilon w^\varepsilon \phi \, dx \to \int_{\Omega \times Y} v_0(x, y) w(y) \phi(x) \, dx \, dy \quad \forall w \in L_p^2, \quad \forall \phi \in \mathscr{K}(\bar{\Omega}),$$

where $v_0 \in L^2(\Omega; L_p^2)$.

We will also need the following lemma.

LEMMA 2. *Let* $f = (f^i) \in \mathbf{L}_p^2(Y_0)$. *Assume that*

$$\sum_{i=1}^N \int_{Y_0} f^i w^i \, dy = 0 \quad \textit{for all } w = (w^i) \in \mathscr{D}_p(Y_0), \quad \operatorname{div}_y w = 0. \tag{2.9}$$

Then there exists a unique function $q \in H_p^1(Y_0)/\mathbb{C}$ *such that*

$$\frac{\partial q}{\partial y_i} = f^i \quad \textit{for } i = 1, \cdots, N. \tag{2.10}$$

Lemma 2 is well known in its nonperiodic version (whose proof can be seen in [24, pp. 14, 15]). The proof in the present version is quite similar to the latter, by use of Propositions 1.1 and 1.2 from [24, pp. 14, 15] and the above property of $\tilde{Q}_j$ of "primitive preserving with respect to y_j" (see § 2.1).

2.3. Proof of Theorem 1. The proof of Theorem 1 proceeds in four steps:

(i) First, by Lemma 1, (2.7) follows immediately from (2.4). At the same time we have, by weak compactness, $v_\varepsilon \to z_0$ in $L^2(\Omega)$-weak, with $z_0 = \tilde{v}_0$, which proves (2.6).

(ii) From now on, ε denotes the above subsequence. Define $z_\varepsilon^i \in L^2(\Omega)$, $1 \leqq i \leqq N$, such that

$$\int_\Omega z_\varepsilon^i v \, dx = \int_{\Omega_\varepsilon^0} \frac{\partial v_\varepsilon}{\partial x_i} v \, dx \quad \forall v \in L^2(\Omega). \tag{2.11}$$

Again by Lemma 1 we deduce from (2.5) that

$$\int_\Omega z_\varepsilon^i w^\varepsilon \phi \, dx \to \int_{\Omega \times Y_0} z^i(x, y) w(y) \phi(x) \, dx \, dy \tag{2.12}$$

$$\forall w \in L_p^2, \quad \forall \phi \in \mathscr{K}(\bar{\Omega}),$$

where $z^i \in L^2(\Omega; L_p^2)$ with $z^i(x, \cdot) = 0$ outside Y_0 for almost all x in Ω.

Next, we show that the function v_0 in (2.7) can be decomposed as stated above. Take in (2.11) test functions of the form $v = \varepsilon w^\varepsilon \phi$ with $w \in \mathcal{D}_p(Y_0)$, $\phi \in \mathcal{D}(\Omega)$. We have easily that

$$\varepsilon \int_\Omega z_\varepsilon^i w^\varepsilon \phi \, dx = -\int_\Omega v_\varepsilon \left(\frac{\partial w}{\partial y_i}\right)^\varepsilon \phi \, dx - \varepsilon \int_\Omega v_\varepsilon w^\varepsilon \frac{\partial \phi}{\partial x_i} \, dx.$$

Hence, letting $\varepsilon \downarrow 0$ and recalling (2.7) and (2.12) gives

$$\int_{\Omega \times Y} v_0(x, y) \frac{\partial w}{\partial y_i}(y) \phi(x) \, dx \, dy = 0 \quad \forall w \in \mathcal{D}_p(Y_0), \quad \forall \phi \in \mathcal{D}(\Omega).$$

Therefore, for almost all $x \in \Omega$ we have $\partial v_0 / \partial y_i \, (x, \cdot) = 0$ in Y_0, $i = 1, \cdots, N$. That is, in sum, there exists $u \in L^2(\Omega)$ such that $v_0(x, y) = u(x)$ for almost all $y \in Y_0$. The desired decomposition of v_0 follows at once from defining u_r as $u_r = v_0 - u$, provided, of course, that $u \in H^1(\Omega)$.

To end this step, let us show that $u \in H^1(\Omega)$. Choosing in (2.11) test functions of the form $v = w^\varepsilon \phi$, $\phi \in \mathcal{D}(\Omega)$ and $w \in \mathcal{D}_p(Y_0)$ with $\partial w / \partial y_i = 0$ for some fixed i (note that by (2.2) such w's exist), we easily arrive at

$$\int_\Omega z_\varepsilon^i w^\varepsilon \phi \, dx = -\int_\Omega v_\varepsilon w^\varepsilon \frac{\partial \phi}{\partial x_i} \, dx.$$

By (2.7) (where henceforth $v_0 = u + u_r$) it follows that

$$\int_\Omega \left(\int_{Y_0} z^i(x, y) w(y) \, dy \right) \phi(x) \, dx = \left(-\int_\Omega u \frac{\partial \phi}{\partial x_i} \, dx \right) \int_{Y_0} w(y) \, dy.$$

Hence, choosing w so that $\int_{Y_0} w \, dy = 1$, we have

$$\left\langle \frac{\partial u}{\partial x_i}, \phi \right\rangle = \int_\Omega \left(\int_{Y_0} z^i(x, y) w(y) \, dy \right) \phi(x) \, dx \quad \forall \phi \in \mathcal{D}(\Omega),$$

$1 \leqq i \leqq N$ ($\langle , \rangle$ denoting the usual duality between $\mathcal{D}'$ and $\mathcal{D}$), which shows that $u \in H^1(\Omega)$.

(iii) In what follows, ε designates the subsequence extracted in step (ii). The next point is to check that there exists u_1 in $L^2(\Omega; H_p^1(Y_0)/\mathbb{C})$ such that

$$z^i(x, y) = \frac{\partial u}{\partial x_i}(x) + \frac{\partial u_1}{\partial y_i}(x, y) \quad \text{for } i = 1, \cdots, N. \tag{2.13}$$

But, based on Lemma 2, it is easy to see that this proceeds exactly as in [16, proof of Thm. 3].

(iv) The last point is to show that $v_\varepsilon \in H_0^1(\Omega)$ implies $u \in H_0^1(\Omega)$. For arbitrarily fixed i, let $w \in \mathcal{D}_p(Y_0)$ with $\partial w / \partial y_i = 0$ and $\int_{Y_0} w \, dy = 1$. Let $\phi \in \mathcal{D}(\bar{\Omega})$ (the subspace of $\mathcal{H}(\bar{\Omega})$ made up of C^∞ functions). Take $v = w^\varepsilon \phi$ in (2.11). Noting that v vanishes outside Ω_ε^0, an elementary integration by parts on the right of (2.11) yields

$$\int_\Omega z_\varepsilon^i w^\varepsilon \phi \, dx = -\int_\Omega v_\varepsilon w^\varepsilon \frac{\partial \phi}{\partial x_i} \, dx,$$

where ε designates the subsequence involved in step (iii). Letting $\varepsilon \downarrow 0$ and using (2.12) and (2.13) leads to

$$\int_\Omega \frac{\partial u}{\partial x_i} \phi \, dx = -\int_\Omega u \frac{\partial \phi}{\partial x_i} \, dx \quad (1 \leqq i \leqq N) \quad \forall \phi \in \mathcal{D}(\bar{\Omega}),$$

which shows that $u = 0$ on $\partial\Omega$. The proof is complete. □

Remark 4. The subsequence ε involved in Theorem 1 is precisely the one we extracted in step (ii) above.

Remark 5. Assume that $\|v_\varepsilon\|_{H^1(\Omega)} \leqq c$ for all ε. Then by extraction of a suitable subsequence we have [16]

$$v_\varepsilon \to u \quad \text{in } H^1(\Omega)\text{-weak},$$

$$\int_\Omega \frac{\partial v_\varepsilon}{\partial x_i} w^\varepsilon \phi \, dx \to \int_{\Omega \times Y} \left(\frac{\partial u}{\partial x_i}(x) + \frac{\partial u_1}{\partial y_i}(x, y) \right) w(y) \phi(x) \, dx \, dy$$

$$\forall w \in L^2_p, \quad \forall \phi \in \mathscr{K}(\bar{\Omega}),$$

where $u_1 \in L^2(\Omega; H^1_p(Y)/\mathbb{C})$.

3. Analysis of the problem. As mentioned in § 1, two distinct local structures (relative to Figs. 1 and 2, respectively) are considered in this work, so that our study actually deals with two homogenization problems associated with the variational problem (1.10) (or (1.12)). Nevertheless, both models are studied simultaneously, with mention being made of the specific results in each case.

3.1. Preliminary results. We start with some estimates. In what follows, c denotes various constants independent of variable quantities (such as ε, $t, \cdots$). First, let $z_\varepsilon(t) = e^{-rt} u_\varepsilon(t)$ (u_ε the solution of (1.10)), where r is a fixed real number. Then taking $v = \partial z_\varepsilon / \partial t(t)$ in (1.10) and using routine estimating techniques leads to

$$\|z_\varepsilon(t)\|^2_{\mathbf{L}^2(\Omega)} + \varepsilon^2 b^\varepsilon(z_\varepsilon(t), z_\varepsilon(t)) + c^\varepsilon(z_\varepsilon(t), z_\varepsilon(t)) \leqq c \int_0^T e^{-rs} \|f(s)\|^2_{\mathbf{L}^2(\Omega)} \, ds$$

for almost all $0 < t < +\infty$, for all ε, and all $T > 0$. Therefore, assuming that r is large enough and using (1.9) gives

$$\|z_\varepsilon(t)\|^2_{\mathbf{L}^2(\Omega)} + \varepsilon^2 b^\varepsilon(z_\varepsilon(t), z_\varepsilon(t)) + c^\varepsilon(z_\varepsilon(t), z_\varepsilon(t)) \leqq c$$

for almost all $0 < t < +\infty$ and for all ε.

Hence, referring to the explicit definition of the form c^ε (Remark 1) and using (1.6) we obtain

$$\|z_\varepsilon(t)\|^2_{\mathbf{L}^2(\Omega)} + \|\operatorname{div} z_\varepsilon(t)\|^2_{L^2(\Omega)} \leqq c \tag{3.1}$$

for all ε and almost all $t > 0$.

We now investigate the consequences of the above estimate. We introduce the well-known Hilbert space (see, e.g., [24])

$$E_0(\Omega) = \{w;\ w \in \mathbf{L}^2(\Omega),\ \operatorname{div} w \in L^2(\Omega) \text{ and } w.n = 0 \text{ on } \partial\Omega\}$$

(n representing the outer unit normal to $\partial\Omega$), equipped with the norm $\|w\|_{E(\Omega)} = (\|w\|^2_{\mathbf{L}^2(\Omega)} + \|\operatorname{div} w\|^2_{L^2(\Omega)})^{1/2}$.

Clearly we have, for almost all $t > 0$, $z_\varepsilon(t) \in E_0(\Omega)$. Moreover by virtue of (3.1) the sequence (z_ε) remains bounded in the space $L^\infty(0, +\infty; E_0(\Omega))$. It follows, by extraction of a subsequence:

$$z_\varepsilon \to z_0 \quad \text{in } L^\infty(0, +\infty; E_0(\Omega))\text{-weak star}. \tag{3.2}$$

On letting $u_0(t) = z_0(t)\, e^{rt}$ we deduce that

$$\hat{u}_\varepsilon(\lambda) \to \hat{u}_0(\lambda) \quad \text{in } E_0(\Omega)\text{-weak for any } \lambda \in \mathbb{C}, \quad \operatorname{Re} \lambda > \lambda_0 > r, \tag{3.3}$$

$$u_\varepsilon \to u_0 \quad \text{in } L^\infty(0, T; E_0(\Omega))\text{-weak star for any } T > 0. \tag{3.4}$$

Now, choosing in (1.12) $v = \hat{u}_\varepsilon(\lambda)$ (with $\mathrm{Re}\,\lambda > \lambda_0 > 0$) yields

$$c^\varepsilon(\hat{u}_\varepsilon, \hat{u}_\varepsilon) \leqq c \quad \forall \varepsilon, \tag{3.5}$$

$$\varepsilon^2 b^\varepsilon(\hat{u}_\varepsilon, \hat{u}_\varepsilon) \leqq c \quad \forall \varepsilon, \tag{3.6}$$

where, for the sake of simplicity, we write $\hat{u}_\varepsilon$ in place of $\hat{u}_\varepsilon(\lambda)$. Thanks to the property $b^\varepsilon(v, v) + c^\varepsilon(v, v) \geqq c\|v\|^2_{\mathbf{H}^1_0(\Omega)}$ for all $v \in \mathbf{H}^1_0(\Omega)$ and all $\varepsilon > 0$, it follows from (3.5) and (3.6) that

$$\varepsilon \|\hat{u}_\varepsilon\|_{\mathbf{H}^1(\Omega)} \leqq c \quad \forall \varepsilon. \tag{3.7}$$

On the other hand, from (3.3) we obtain

$$\|\mathrm{div}\ \hat{u}_\varepsilon\|_{L^2(\Omega)} \leqq c \quad \forall \varepsilon. \tag{3.8}$$

To complete these estimates, let us give a crucial lemma.

LEMMA 3. *There is some ε_0 such that*

$$\sum_{i,j=1}^{3} \int_{\Omega^s_\varepsilon} \left| \frac{\partial \hat{u}^i_\varepsilon}{\partial x_j} \right|^2 dx \leqq c \quad \textit{for all } 0 < \varepsilon < \varepsilon_0.$$

Proof. The properties (1.6) and (3.5) give $\int_{\Omega^s_\varepsilon} E_{ij}(\hat{u}_\varepsilon)\overline{E_{ij}(\hat{u}_\varepsilon)}\, dx \leqq c$ (the summation convention is used), and the lemma follows immediately by Theorem A (in the Appendix) combined with Korn's inequality [10]. □

We are now able to obtain some specific convergence results needed in the sequel. Unless otherwise specified, the summation convention is used throughout the rest of this section and also in § 4. Also, the complex number λ (associated with the Laplace transform) is fixed (with $\mathrm{Re}\,\lambda > \lambda_0 > 0$, λ_0 large enough) and therefore is omitted.

LEMMA 4. *We can extract a subsequence ε such that*

$$\int_\Omega \hat{u}^k_\varepsilon \psi^\varepsilon \phi\, dx \to \int_{\Omega \times Y} w^k_0(x, y)\psi(y)\phi(x)\, dx\, dy, \qquad 1 \leqq k \leqq 3, \tag{3.9}$$

and

$$\int_\Omega \varepsilon \frac{\partial \hat{u}^k_\varepsilon}{\partial x_h} \psi^\varepsilon \phi\, dx \to \int_{\Omega \times Y} \frac{\partial w^k_0}{\partial y_h}(x, y)\psi(y)\phi(x)\, dx\, dy, \qquad 1 \leqq k,\ h \leqq 3, \tag{3.10}$$

$$\forall \psi \in L^2_p, \quad \phi \in \mathscr{K}(\bar{\Omega}),$$

where

$$w_0 = (w^k_0) \in L^2(\Omega; \mathbf{H}^1_p), \tag{3.11}$$

$$\mathrm{div}_y\, w_0 = 0. \tag{3.12}$$

Moreover,

$$\hat{u}_0(x) = \int_Y w_0(x, y)\, dy \qquad (u_0 \textit{ is the function in } (3.4)). \tag{3.13}$$

Proof. Property (3.9) is straightforward by Lemma 1, the sequence $(\hat{u}^k_\varepsilon)_{\varepsilon > 0}$ being bounded in $\mathbf{L}^2(\Omega)$. Property (3.10) results from (3.7) by application of the results in Remark 5 and use of (3.9). Equation (3.12) derives immediately from (3.8) and (3.10). As for (3.13), it suffices to use (3.3) and (3.9). □

LEMMA 5. *Let ε be the subsequence involved in Lemma 4. Then, as $\varepsilon \downarrow 0$, we have*

$$\varepsilon^2 \int_\Omega b^\varepsilon_{ijkh} \frac{\partial \hat{u}^k_\varepsilon}{\partial x_h} \frac{\partial (w^i)^\varepsilon}{\partial x_j} \phi \, dx \to \mu_{ijkh} \int_{\Omega \times Y_f} \frac{\partial w^k_0}{\partial y_h}(x, y) \frac{\partial w^i}{\partial y_j}(y) \phi(x) \, dx \, dy$$

$$\forall w = (w^i) \in \mathbf{H}^1_p, \quad \forall \phi \in \mathscr{K}(\bar{\Omega}),$$

where $\mu_{ijkh} = \mu(\delta_{ik}\delta_{jh} + \delta_{jk}\delta_{ih})$.

Proof. The desired property follows at once from (3.10) by the choice $\psi = b_{ijkh}(\partial w^i / \partial y_j)$ (summation over $1 \leqq i, j \leqq 3$) with $w \in \mathbf{H}^1_p$. □

LEMMA 6. *A subsequence can be extracted from the one in Lemma 4 such that*

$$\int_{\Omega^s_\varepsilon} \frac{\partial \hat{u}^k_\varepsilon}{\partial x_h} \psi^\varepsilon \phi \, dx \to \int_{\Omega \times Y_s} \left[\frac{\partial u^k}{\partial x_h}(x) + \frac{\partial u^k_1}{\partial y_h}(x, y) \right] \psi(y) \phi(x) \, dx \, dy, \tag{3.14}$$

$$1 \leqq k, h \leqq 3, \quad \forall \psi \in L^2_p, \quad \forall \phi \in \mathscr{K}(\bar{\Omega}),$$

where $u = (u^k) \in \mathbf{H}^1_0(\Omega)$, $u_1 = (u^k_1) \in L^2(\Omega; \mathbf{H}^1_p(Y_s)/\mathbb{C}^3)$.

Moreover the limit w_0 in Lemma 4 decomposes as follows:

$$w_0(x, y) = u(x) + u_r(x, y) \tag{3.15}$$

with $u_r \in L^2(\Omega; \mathbf{H}^1_p)$, $u_r(x, y) = 0$ for $y \in Y_s$ (x fixed), $\operatorname{div}_y u_r = 0$.

The above results follow immediately from Lemma 3 by application of Theorem 1 and use of Lemma 4.

We end these preliminaries with the study of the behaviour of the pressure p_ε in the fluid part Ω^f_ε. We put

$$\hat{p}_\varepsilon = -\gamma \operatorname{div} \hat{u}_\varepsilon \quad \text{in } \Omega^f_\varepsilon \quad (\gamma \text{ defined in § 1}). \tag{3.16}$$

We have $\hat{p}_\varepsilon \in L^2(\Omega^f_\varepsilon)$ with $\|\hat{p}_\varepsilon\|_{L^2(\Omega^f_\varepsilon)} \leqq c$ for all $\varepsilon > 0$. Next, we define $D^{kh}_\varepsilon \in L^2(\Omega)$ ($1 \leqq k, h \leqq 3$) such that

$$\int_\Omega D^{kh}_\varepsilon v \, dx = \int_{\Omega^s_\varepsilon} \frac{\partial \hat{u}^k_\varepsilon}{\partial x_h} v \, dx - \frac{\delta_{kh}}{3\gamma} \int_{\Omega^f_\varepsilon} \hat{p}_\varepsilon v \, dx \quad \forall v \in L^2(\Omega) \tag{3.17}$$

(δ_{kh} is the Kronecker symbol). The sequence $(D^{kh}_\varepsilon)_{\varepsilon > 0}$ remains in a bounded set of $L^2(\Omega)$. Denoting by D^{kh} its weak limit in the sense of Lemma 1, it is easy to verify (by Lemma 6) that

$$D^{kh}(x, y) = \frac{\partial u^k}{\partial x_h}(x) + \frac{\partial u^k_1}{\partial y_h}(x, y) \quad \text{for } (x, y) \in \Omega \times \tilde{Y}_s.$$

Finally, on letting $p_0(x, y) = -\gamma D^{kk}(x, y)$ (summation over $1 \leqq k \leqq 3$) for $(x, y) \in \Omega \times \tilde{Y}_f$, we easily obtain (by combination of the preceding relation and (3.15)) the following lemma.

LEMMA 7. *As $\varepsilon \downarrow 0$ (ε a subsequence from the one in Lemma 6) we have for all $\psi \in L^2_p$ and all $\phi \in \mathscr{K}(\bar{\Omega})$:*

$$\int_{\Omega^f_\varepsilon} \hat{p}_\varepsilon \psi^\varepsilon \phi \, dx \to \int_{\Omega \times Y_f} p_0(x, y) \psi(y) \phi(x) \, dx \, dy, \quad p_0 \in L^2(\Omega; L^2_p(Y_f)). \tag{3.18}$$

Moreover,

$$\int_{Y_s} \operatorname{div}_y u_1(\cdot, y) \, dy = |Y_f| \operatorname{div} u + \operatorname{div} \int_{Y_f} u_r(\cdot, y) \, dy + \frac{1}{\gamma} \int_{Y_f} p_0(\cdot, y) \, dy (|Y_f| = \text{meas } Y_f). \tag{3.19}$$

Remark 6. In the sequel, ε represents the subsequence involved in Lemma 7. Observe that Lemmas 4–7 hold simultaneously for that subsequence.

3.2. Derivation of the local problems. The aim in this section is to derive the problems for $u_1(x, y)$ and $u_r(x, y)$, respectively.

The local problem for u_1. We adapt the idea in [16, § 6]. We choose in (1.12) the v's of the form $v = \varepsilon w^\varepsilon \phi$; $w \in \mathbf{H}^1_p$, $\phi \in \mathscr{D}(\Omega)$. Next we pass to the limit by Lemma 6 (with $\psi = a_{ijkh}(\overline{\partial w^i / \partial y_j})$) and by Lemma 7 (with $\psi = \overline{\operatorname{div}_y w}$) and we obtain the local problem for u_1:

$$\int_{Y_s} a_{ijkh}\left[\frac{\partial u^k}{\partial x_h}(x) + \frac{\partial u_1^k}{\partial y_h}(x, y)\right]\overline{\frac{\partial w^i}{\partial y_j}}(y)\, dy - \int_{Y_f} p_0(x, y)\,\overline{\operatorname{div}_y w(y)}\, dy = 0 \quad \forall w \in \mathbf{H}^1_p. \tag{3.20}$$

Immediately we have that p_0 does not depend on y. In other words, $p_0 \in L^2(\Omega)$. With this in mind, an elementary operation and use of the extension operator T_p (see the Appendix) show that we may replace, in (3.20), $\mathbf{H}^1_p$ by $\mathbf{H}^1_p(Y_s)/\mathbb{C}^3$. Hence, the local problem for u_1 takes the more precise form (where x is fixed):

$$\begin{aligned} &u_1(x, \cdot) \in \mathbf{H}^1_p(Y_s)/\mathbb{C}^3, \\ &a(u_1(x, \cdot), w) = -\frac{\partial u^k}{\partial x_h}(x)\int_{Y_s} a_{ijkh}\overline{\frac{\partial w^i}{\partial y_j}}\, dy - p_0(x)\int_{Y_s}\overline{\operatorname{div}_y w}\, dy \\ &\forall w \in \mathbf{H}^1_p(Y_s)/\mathbb{C}^3, \end{aligned} \tag{3.21}$$

where $a(,)$ represents the real sesquilinear form given by

$$a(v, w) = \int_{Y_s} a_{ijkh}\frac{\partial v^k}{\partial y_h}\overline{\frac{\partial w^i}{\partial y_j}}\, dy. \tag{3.22}$$

It is easy to see that $u_1(x, \cdot)$ is the unique solution of the variational problem (3.21): It suffices to check that the form $a(,)$ is coercive on $\mathbf{H}^1_p(Y_s)/\mathbb{C}^3$, i.e., there exists $c > 0$ such that

$$a(w, w) \geqq c\|w\|^2_{\mathbf{H}^1(Y_s)/\mathbb{C}^3} \quad \forall w \in \mathbf{H}^1_p(Y_s)/\mathbb{C}^3. \tag{3.23}$$

But this results from (1.5), (1.6), and Theorem B (in the Appendix) combined with the following elementary inequality:

$$\int_Y e_{ij}(w)\overline{e_{ij}(w)}\, dy \geqq \frac{1}{2}\int_Y \frac{\partial w^i}{\partial y_j}\overline{\frac{\partial w^i}{\partial y_j}}\, dy \quad \forall w \in \mathbf{H}^1_p.$$

The next point is to calculate u_1 in terms of u and p_0. To this end, we define vector functions (independent of x) χ, $\chi^j_i \in \mathbf{H}^1_p(Y_s)/\mathbb{C}^3$ $(1 \leqq i, j \leqq 3)$ by

$$a(\chi, w) = \int_{Y_s}\overline{\operatorname{div}_y w}\, dy \quad \forall w \in \mathbf{H}^1_p(Y_s)/\mathbb{C}^3 \tag{3.24}$$

and

$$a(\chi^j_i, w) = \int_{Y_s} a_{ijkh}\overline{\frac{\partial w^k}{\partial y_h}}\, dy \quad \forall w \in \mathbf{H}^1_p(Y_s)/\mathbb{C}^3, \tag{3.25}$$

respectively. By virtue of (3.23) the preceding functions are uniquely determined. Moreover, they are actually real-valued vector functions, i.e., $\chi(y)$, $\chi_i^j(y) \in \mathbb{R}^3$.

Then, by virtue of uniqueness in (3.21) we have Lemma 8.

LEMMA 8. *The function u_1 is given by*

$$u_1(x, y) = -\frac{\partial u^k}{\partial x_h}(x)\chi_k^h(y) - p_0(x)\chi(y). \tag{3.26}$$

In the sequel we will put

$$\beta_{ij} = -\int_{Y_s} \operatorname{div}_y \chi_i^j \, dy, \quad \beta = \int_{Y_s} \operatorname{div}_y \chi \, dy \quad \text{(note that } \beta = a(\chi, \chi) > 0). \tag{3.27}$$

For further needs it is useful to see that (3.26) permits us to express $p_0(x)$ in terms of $u(x)$ and $\tilde{u}_r(x) = \int_{Y_f} u_r(x, y)\, dy$. Indeed, substitution of (3.26) into (3.19) yields

$$\delta^{-1} p_0 = \beta_{kh} \frac{\partial u^k}{\partial x_h} - \Pi \operatorname{div} u - \operatorname{div} \tilde{u}_r, \tag{3.28}$$

where

$$\delta = \left(\frac{\Pi}{\gamma} + \beta\right)^{-1}, \qquad \Pi = \frac{|Y_f|}{|Y|} \quad \text{(porosity)} \tag{3.29}$$

(we recall the general notation $|\mathcal{O}|$ = measure of $\mathcal{O} \subset \mathbb{R}^3$).

The local problem for u_r. In what follows, we set

$$W = \{w \in \mathbf{H}_p^1;\ w = 0 \text{ on } Y_s \text{ and } \operatorname{div}_y w = 0\}.$$

W is a closed vector subspace of $\mathbf{H}_p^1$ and we have $u_r \in L^2(\Omega; W)$ (see Lemma 6). Now take in (1.12) the v's of the form $v = w^\varepsilon \phi$, $w \in W$, and $\phi \in \mathcal{D}(\Omega)$. Then, letting $\varepsilon \downarrow 0$ and using Lemmas 4, 5, and 7 leads to the local problem for u_r:

$$\begin{aligned} &\lambda^2 \rho^f \int_{Y_f} u_r^i(x, \cdot) w^i \, dy + \lambda\mu \int_{Y_f} \frac{\partial u_r^i}{\partial y_j}(x, \cdot) \frac{\partial w^i}{\partial y_j} \, dy \\ &= \left(\hat{f}^i(x) - \lambda^2 \rho^f u^i(x) - \frac{\partial p_0}{\partial x_i}(x)\right) \int_{Y_f} w^i \, dy \quad \forall w \in W. \end{aligned} \tag{3.30}$$

Let us take the opportunity to give a property in connection with the local topological structure.

LEMMA 9. *For the model represented by Fig. 1 we have $u_r = 0$.*

Proof. By an elementary calculation using the Stokes formula we have

$$\int_{Y_f} w^i \, dy = 0 \quad (i = 1, 2, 3) \quad \forall w = (w^i) \in W, \tag{3.31}$$

and the desired property follows by substitution into (3.30). □

4. The homogenized problem and the convergence theorem. Our goal in this section is to derive the boundary value problem for w_0, the weak limit (in the sense of Lemma 1) of $\hat{u}_\varepsilon$, next, to prove its uniqueness, and finally, to establish the convergence of u_ε to u_0 (thus far, this holds only for a subsequence; see (3.4)).

4.1. Derivation of the macroscopic equation. The procedure (see [16]) consists of fixing v in (1.12), with $v \in \mathscr{D}(\Omega)$, and concentrating on passing to the limit as $\varepsilon \downarrow 0$. With Lemmas 1, 4, 6, and 7, we are in a position to do so. We easily arrive at

$$\begin{aligned} &\lambda^2 \int_\Omega (\tilde{\rho} u^i + \rho^f \tilde{u}_r^i) v^i \, dx - \Pi \int_\Omega p_0 \operatorname{div} v \, dx \\ &+ \int_{\Omega \times Y_s} a_{ijkh} \left(\frac{\partial u^k}{\partial x_h} + \frac{\partial u_1^k}{\partial y_h} \right) \frac{\partial v^i}{\partial x_j} \, dx \, dy \\ &= \int_\Omega \hat{f}^i v^i \, dx \quad \forall v \in \mathscr{D}(\Omega), \end{aligned} \tag{4.1}$$

where the following notation is used

$$\tilde{w} = \int_Y w(y) \, dy \quad \text{for } w \in L^2_p. \tag{4.2}$$

We now write (4.1) in a more appropriate form. For $1 \leqq i, j \leqq 3$, we introduce:

$$p_i^j \text{ the vector function whose components are } y_j \delta_{ik} \ (k = 1, 2, 3), \tag{4.3}$$

$$q_{ijkh} = a(\chi_i^j - p_i^j, \chi_k^h - p_k^h), \tag{4.4}$$

where $a(,)$ (respectively, χ_i^j) is defined in (3.22) (respectively, (3.25)). Then, following [2], an elementary calculation gives

$$\int_{Y_s} a_{ijkh} \left(\frac{\partial u^k}{\partial x_h} + \frac{\partial u_1^k}{\partial y_h} \right) dy = q_{ijkh} \frac{\partial u^k}{\partial x_h} + \beta_{ij} p_0 .$$

Substituting this into (4.1) and using the fact that $\mathscr{D}(\Omega)$ is dense in $\mathbf{H}_0^1(\Omega)$ we obtain the so-called macroscopic equation:

$$\begin{aligned} &\lambda^2 \int_\Omega (\tilde{\rho} u^i + \rho^f \tilde{u}_r^i) v^i \, dx + \int_\Omega q_{ijkh} \frac{\partial u^k}{\partial x_h} \frac{\partial v^i}{\partial x_j} \, dx \\ &+ \delta \int_\Omega \left(\beta_{kh} \frac{\partial u^k}{\partial x_h} - \Pi \operatorname{div} u - \operatorname{div} \tilde{u}_r \right) \left(\beta_{ij} \frac{\partial v^i}{\partial x_j} - \Pi \operatorname{div} v \right) dx \\ &= \int_\Omega \hat{f}^i v^i \, dx \quad \forall v \in \mathbf{H}_0^1(\Omega). \end{aligned} \tag{4.5}$$

In what follows, it is crucial to note that the coefficients q_{ijkh} are real and they possess the classical properties of symmetry and ellipticity:

$$q_{ijkh} = q_{jikh} = q_{ijhk} = q_{khij}, \tag{4.6}$$

$$q_{ijkh} \xi_{kh} \xi_{ij} \geqq c \xi_{ij} \xi_{ij} \quad (c > 0) \quad \forall \xi_{ij} \in \mathbb{R}, \quad \xi_{ij} = \xi_{ji} \quad (1 \leqq i, j \leqq 3). \tag{4.7}$$

The latter property is easily obtained after the classical fashion of [2].

4.2. The limit problem for the model represented by Fig. 1. Let us first resume the preceding analysis. In § 3.1 we found a function $u_0 \in L^\infty_{\text{loc}}(\mathbb{R}; E_0(\Omega))$ with $u_0(t) = 0$ for $t < 0$, and a subsequence ε (from the sequence involved in (1.10)) such that (3.3) and (3.4) hold as $\varepsilon \downarrow 0$. In § 3.2 we found that for each λ there exists a function $u(\lambda) \in \mathbf{H}_0^1(\Omega)$ such that $\hat{u}_0(\lambda) = u(\lambda)$. So, for the present model, $\hat{u}_0$ is actually a function of λ with values in $\mathbf{H}_0^1(\Omega)$.

Now, we deduce from (4.5) and the above results (and, in particular, Lemma 9) that for any λ ($\operatorname{Re}\lambda > \lambda_0$) $\hat{u}_0(\lambda)$ is a solution of the following variational problem (where, for simplicity, λ is omitted):

(4.8) Find $\hat{u}_0 \in \mathbf{H}_0^1(\Omega)$ such that

$$\lambda^2 \int_\Omega \tilde{\rho} \hat{u}_0^i \bar{v}^i \, dx + Q(\hat{u}_0, v) = \int_\Omega \hat{f}^i \overline{v^i} \, dx \quad \forall v \in \mathbf{H}_0^1(\Omega),$$

where

$$Q(v, w) = \int_\Omega q_{ijkh} \frac{\partial v^k}{\partial x_h} \overline{\frac{\partial w^i}{\partial x_j}} \, dx + \delta \int_\Omega \left(\beta_{kh} \frac{\partial v^k}{\partial x_h} - \Pi \operatorname{div} v \right) \left(\beta_{ij} \overline{\frac{\partial w^i}{\partial x_j}} - \Pi \overline{\operatorname{div} w} \right) dx.$$

The sesquilinear form Q is coercive on $\mathbf{H}_0^1(\Omega)$ by virtue of (4.7), so that $\hat{u}_0(\lambda)$ is uniquely determined by problem (4.8). Therefore, the sequence ε in (3.3) and (3.4) may be replaced by the whole sequence from which it was extracted.

Finally, here is our convergence theorem for the model represented by Fig. 1.

THEOREM 2. *Let u_ε be the solution of* (1.10) *for the model represented by Fig.* 1. *Then for any fixed $T > 0$ we have $u_\varepsilon \to u_0$ in $L^\infty(0, T; E_0(\Omega))$-weak star as $\varepsilon \downarrow 0$, where u_0 is the unique distribution in $\mathscr{D}'(\mathbb{R}; \mathbf{H}_0^1(\Omega)) \cap L^\infty_{\mathrm{loc}}(\mathbb{R}; E_0(\Omega))$ with support in $[0, +\infty[$, having a Laplace transform $\hat{u}_0(\lambda)$, which is (for $\operatorname{Re}\lambda > \lambda_0 > 0$, λ_0 large enough) the unique solution of* (4.8).

Proof. We need only show that $\hat{u}_0$ is the Laplace transform of a distribution in $\mathscr{D}'(\mathbb{R}; \mathbf{H}_0^1(\Omega))$ with support in $[0, +\infty[$. This will be the fact if we verify that:

(i) $\hat{u}_0$ *is holomorphic of* $\{\lambda \in \mathbb{C}; \operatorname{Re}\lambda > \lambda_0\}$ into $\mathbf{H}_0^1(\Omega)$, and

(ii) $\|\hat{u}_0\|_{\mathbf{H}_0^1(\Omega)}$ is bounded by a polynomial in $|\lambda|$.

Part (i) is obtained by a classical criterion in [11, p. 365] combined with an elementary operation in (4.8) that takes into account two facts: on the one hand, $\hat{u}_0$ and $\hat{f}$ are holomorphic with values in $\mathbf{L}^2(\Omega)$; on the other hand, $\mathbf{L}^2(\Omega)$ (identified with its antidual) is dense in $E_0'(\Omega)$ (antidual of $E_0(\Omega)$) and also in $\mathbf{H}^{-1}(\Omega)$.

Part (ii) is straightforward by taking $v = \hat{u}_0(\lambda)$ in (4.8) and effecting a routine estimate after the fashion of, e.g., [18, p. 162]. The proof is complete. □

Remark 7. The evolution homogenized problem is as follows:

(4.9) Find u_0, a function of $0 < t < +\infty$ with values in $\mathbf{H}_0^1(\Omega)$ such that

$$\tilde{\rho} \int_\Omega \frac{\partial^2 u_0^i}{\partial t^2} \overline{v^i} \, dx + Q(u_0, v) = \int_\Omega f^i \overline{v^i} \, dx \quad \forall v \in \mathbf{H}_0^1(\Omega),$$

$$u_0(0) = \frac{\partial u_0}{\partial t}(0) = 0.$$

Existence and uniqueness in (4.9) can be proved, for example, by use of the theory of semigroups (for sufficiently regular f). Furthermore, (4.8) proves to be the "Laplace transform" of (4.9), so that the limit u_0 in Theorem 2 can be directly characterized by (4.9).

4.3. The limit problem for the model represented by Fig. 2. In the present situation relative to Fig. 2, Lemma 9 is no longer true (indeed, (3.31) does not hold): we have, in general, $u_r \neq 0$. Therefore, in opposition to the preceding situation, (4.5) is not by itself the desired limit problem. As we will see in the sequel, the limit (or homogenized)

problem is obtained by combining (3.30) and (4.5). In what follows we use the notation $\tilde{w}(x) = \int_Y w(x, y)\, dy$, $w \in L^2(\Omega; L^2_p)$.

We first put (3.30) in a suitable form: After substituting (3.28) into (3.30) we take test functions w of the form $w = v_r(x, \cdot)$ for fixed x, $v_r \in \mathcal{D}(\Omega; W)$, and we integrate over Ω to obtain

$$\begin{aligned} &\lambda^2 \rho^f \int_{\Omega \times Y_f} (u_r^i + u^i) v_r^i \, dx\, dy + \lambda \mu \int_{\Omega \times Y_f} \frac{\partial u_r^i}{\partial y_j} \frac{\partial v_r^i}{\partial y_j} \, dx\, dy \\ &\quad + \delta \int_\Omega \left(\beta_{kh} \frac{\partial u^k}{\partial x_h} - \Pi \operatorname{div} u - \operatorname{div} \tilde{u}_r \right) (-\operatorname{div} \tilde{v}_r) \, dx \\ &\quad = \int_\Omega \hat{f}^i \tilde{v}_r^i \, dx \quad \forall v_r \in \mathcal{D}(\Omega; W). \end{aligned} \tag{4.10}$$

Now, observe that $\tilde{u}_r \in E_0(\Omega)$. This suggests the introduction of the following space:

$$H(\Omega; W) = \{w;\ w \in L^2(\Omega; W), \operatorname{div} \tilde{w} \in L^2(\Omega) \text{ and } \tilde{w} \in E_0(\Omega)\}.$$

We provide $H(\Omega; W)$ with the norm

$$\|w\|_H = (\|w\|^2_{L^2(\Omega; W)} + \|\operatorname{div} \tilde{w}\|^2_{L^2(\Omega)})^{1/2},$$

which makes it a Hilbert space.

We have $u_r \in H(\Omega; W)$, and our intention is to extend (4.10) to all v_r in $H(\Omega; W)$. For this purpose we need the following lemma, whose proof is quite an adaptation of the procedure leading to the analogous result for $\mathcal{D}(\Omega)$ and $E_0(\Omega)$ (see [24, p. 12]).

LEMMA 10. *$\mathcal{D}(\Omega; W)$ is dense in the space $H(\Omega; W)$.*

So, thanks to the above lemma we can replace in (4.10) the set $\mathcal{D}(\Omega; W)$ by the space $H(\Omega; W)$.

At the present time, let us consider (4.5). Observing that the first integral of the left-hand side can take the form

$$(1 - \Pi)\rho^s \int_\Omega u^i v^i \, dx + \rho^f \int_{\Omega \times Y_f} (u^i + u_r^i) v^i \, dx\, dy,$$

we finally combine (4.5) with (4.10) (where $\mathcal{D}(\Omega; W)$ is replaced by $H(\Omega; W)$) to obtain

$$\begin{aligned} &\lambda^2 \rho^f \int_{\Omega \times Y_f} (u^i + u_r^i)(v^i + v_r^i) \, dx\, dy + \lambda^2 (1 - \Pi)\rho^s \int_\Omega u^i v^i \, dx \\ &\quad + \int_\Omega q_{ijkh} \frac{\partial u^k}{\partial x_h} \frac{\partial v^i}{\partial x_j} \, dx + \lambda \mu \int_{\Omega \times Y_f} \frac{\partial u_r^i}{\partial y_j} \frac{\partial v_r^i}{\partial y_j} \, dx\, dy \\ &\quad + \delta \int_\Omega \left(\beta_{kh} \frac{\partial u^k}{\partial x_h} - \Pi \operatorname{div} u - \operatorname{div} \tilde{u}_r \right) \left(\beta_{ij} \frac{\partial v^i}{\partial x_j} - \Pi \operatorname{div} v - \operatorname{div} \tilde{v}_r \right) dx \\ &\quad = \int_\Omega \hat{f} i (v^i + \tilde{v}_r^i) \, dx \quad \forall v \in \mathbf{H}_0^1(\Omega), \quad \forall v_r \in H(\Omega; W). \end{aligned} \tag{4.11}$$

The limit problem for $w_0 = u + u_r$ is thus obtained.

We now prove the uniqueness of w_0. It suffices to verify that the couple (u, u_r) is unique. Or, equivalently, that $\hat{f} = 0$ implies $u = 0$ and $u_r = 0$. So assume $\hat{f} = 0$ in (4.11). Next, choose $v = \bar{u}$ and $v_r = \bar{u}_r$. By division of both sides by λ and taking the real parts, we easily arrive at

$$\int_\Omega q_{ijkh} \frac{\partial u^k}{\partial x_h} \overline{\frac{\partial u^i}{\partial x_j}} \, dx + \mu \int_{\Omega \times Y_f} \frac{\partial u_r^i}{\partial y_j} \overline{\frac{\partial u_r^i}{\partial y_j}} \, dx\, dy \leqq 0 \quad \text{for } \operatorname{Re} \lambda > \lambda_0.$$

Hence, it follows that $\|u\|^2_{\mathbf{H}^1_0(\Omega)} + \|u_r\|^2_{L^2(\Omega;\, W)} \leqq 0$ (use (4.7) for u, and an inequality of Poincaré's type for u_r) and uniqueness in (4.11) follows.

From all that we easily obtain the following convergence theorem.

THEOREM 3. *For each* $\lambda \in \mathbb{C}$, $\operatorname{Re}\lambda > \lambda_0 > 0$ (λ_0 *large enough*), *let* $(u(\lambda), u_r(\lambda)) \in \mathbf{H}^1_0(\Omega) \times H(\Omega; W)$ *be uniquely defined by the variational problem* (4.11). *Let* u_ε *be the solution of* (1.10) *for the periodic local structure represented by Fig.* 2. *Then for each* $T > 0$ *we have, as* $\varepsilon \downarrow 0$:

$$u_\varepsilon \to u_0 \quad \text{in } L^\infty(0, T; E_0(\Omega))\text{-weak star}, \tag{4.12}$$

where u_0 *is the unique function in* $L^\infty_{\mathrm{loc}}(\mathbb{R}; E_0(\Omega))$ *with support in* $[0, +\infty[$ *and having a Laplace transform* $\hat{u}_0(\lambda)$ *given by*

$$\hat{u}_0(x, \lambda) = u(x, \lambda) + \tilde{u}_r(x, \lambda), \qquad \operatorname{Re}\lambda > \lambda_0,$$

with $\tilde{u}_r(x, \lambda) = \int_Y u_r(x, y, \lambda)\, dy$.

Remark 8. In the preceding situation relative to Fig. 1 we have indicated the homogenized problem in the evolution form (4.9). Similarly, it is possible here to give the evolution version of the homogenized problem (4.11).

Let $\mathscr{H} = \mathbf{L}^2(\Omega) \times L^2(\Omega; W)$, $\mathscr{V} = \mathbf{H}^1_0(\Omega) \times H(\Omega; W)$. Elements in $\mathscr{H}$ are denoted by $\mathbf{w} = (w, w_r)$ with $w \in \mathbf{L}^2(\Omega)$, $w_r \in L^2(\Omega; W)$, and the scalar product in $\mathscr{H}$ is denoted by $(\cdot, \cdot)$. Now, for $\mathbf{v} = (v, v_r)$ and $\mathbf{w} = (w, w_r)$ in $\mathscr{V}$ we put

$$\mathbf{Q}(\mathbf{v}, \mathbf{w}) = \int_\Omega q_{ijkh} \frac{\partial v^k}{\partial x_h} \frac{\overline{\partial w^i}}{\partial x_j}\, dx$$
$$+ \delta \int_\Omega \left(\beta_{kh} \frac{\partial v^k}{\partial x_h} - \Pi \operatorname{div} v - \operatorname{div} \tilde{v}_r\right)\left(\beta_{ij} \frac{\overline{\partial w^i}}{\partial x_j} - \Pi\, \overline{\operatorname{div} w} - \overline{\operatorname{div} \tilde{w}_r}\right) dx,$$

$$\mathbf{R}(\mathbf{v}, \mathbf{w}) = \mu \int_{\Omega \times Y_f} \frac{\partial v^i_r}{\partial y_j} \frac{\overline{\partial w^i_r}}{\partial y_j}\, dx\, dy,$$

$$(\mathbf{L}\mathbf{v}, \mathbf{w}) = \rho^f \int_{\Omega \times Y_f} (v^i + v^i_r)(\overline{w^i} + \overline{w^i_r})\, dx\, dy + (1 - \Pi)\rho^s \int_\Omega v^i \overline{w^i}\, dx$$

(this definition being valid for $\mathbf{v}, \mathbf{w} \in \mathscr{H}$).

Next, we introduce $\mathbf{F} \in L^2_{\mathrm{loc}}(0, +\infty; \mathscr{H})$ such that

$$(\mathbf{F}(t), \mathbf{w}) = \int_\Omega f^i(t)(\overline{w^i} + \overline{\tilde{w}^i_r})\, dx \quad \forall \mathbf{w} = (w, w_r) \in \mathscr{H}.$$

Finally, let $\mathbf{z}$ be defined by the following problem:

(4.13) Find $\mathbf{z}$, a function of $0 < t < +\infty$ with values in $\mathscr{V}$, such that

$$(\mathbf{L}\mathbf{z}''(t), \mathbf{v}) + \mathbf{R}(\mathbf{z}'(t), \mathbf{v}) + \mathbf{Q}(\mathbf{z}(t), \mathbf{v}) = (\mathbf{F}(t), \mathbf{v}) \quad \forall \mathbf{v} \in \mathscr{V},$$

$$\mathbf{z}(0) = \mathbf{z}'(0) = 0,$$

where the prime denotes the derivation in the sense of $\mathscr{D}'(0, +\infty; \mathscr{V})$ (see [19] for existence and uniqueness in (4.13)). Then problem (4.11) turns out to be the "Laplace transform" of (4.13), and we have $\hat{\mathbf{z}}(\lambda) = (u(\lambda), u_r(\lambda))$ (the couple in Theorem 3).

Appendix. Extension results.

A1. Statement of the results. With the notation introduced in § 1 (see also § 2), let

$$\Sigma_\varepsilon = \partial\Omega \cap \varepsilon \tilde{Y}_s, \quad V_\varepsilon = \{v \in \mathbf{H}^1(\Omega^s_\varepsilon);\ v = 0 \text{ on } \Sigma_\varepsilon\}, \quad \Omega_1 = \{x \in \mathbb{R}^3;\ d(x, \bar{\Omega}) < 1\},$$

where d designates the Euclidean metric and $\bar{\Omega}$ the closure of Ω in $\mathbb{R}^3$.

THEOREM A. *For each $\varepsilon < \varepsilon_0$ (ε_0 is a suitable constant) there exists an extension operator $T_\varepsilon \in \mathscr{L}(V_\varepsilon, \mathbf{H}_0^1(\Omega_1))$ (i.e., T_ε is continuous linear and $T_\varepsilon u = u$ on Ω_ε^s for all $u \in V_\varepsilon$) such that*

$$\int_{\Omega_1} E_{ij}(T_\varepsilon u)\overline{E_{ij}(T_\varepsilon u)}\, dx \leqq c \int_{\Omega_\varepsilon^s} E_{ij}(u)\overline{E_{ij}(u)}\, dx \quad \forall u \in V_\varepsilon,$$

where the constant c does not depend on ε.

THEOREM B. *There exists an (extension) operator $T_p \in \mathscr{L}(\mathbf{H}_p^1(Y_s), \mathbf{H}_p^1)$ such that $T_p w = w$ almost everywhere in Y_s for all $w \in \mathbf{H}_p^1(Y_s)$ and*

$$\int_Y e_{ij}(T_p w)\overline{e_{ij}(T_p w)}\, dy \leqq c \int_{Y_s} e_{ij}(w)\overline{e_{ij}(w)}\, dy \quad \forall w \in \mathbf{H}_p^1(Y_s).$$

The above theorems are proved in [5] for the periodic structure represented by Fig. 1. We need only study the case of Fig. 2. In fact, technically, the difference between the case of Fig. 1 and that of Fig. 2 lies in the construction of a "local extension," i.e., a suitable (see definition below) extension $p \in \mathscr{L}(H^1(Y_s), H^1(Y))$. Afterwards, the procedure (derivation of T_p and T_ε from p) is quite standard and is not worth repeating here (the interested reader is referred to [5]). Thus, we merely prove the existence of a local extension in the case of Fig. 2 (see Lemma B below).

A2. Local extension. Some notation is needed. Let us fix b, $b > 0$, such that each of the following sets is a cylinder (see Fig. 2):

$$\{y \in Y_f;\ b < y_i < \tfrac{1}{2}\}, \quad \{y \in Y_f;\ -\tfrac{1}{2} < y_i < -b\}, \quad i = 1, 2, 3.$$

Next, fix $a > 0$ (a small enough) and denote by P_1 (respectively, P_2, P_3) the set of points (y_1, r, θ) (respectively, (θ, y_2, r), (r, θ, y_3)) such that $|r - R| < a$, $0 < \theta < 2\pi$, $b < |y_i| < \frac{1}{2}$ for $i = 1$ (respectively, $i = 2$, $i = 3$), (R was introduced in § 1).

On the other hand, let

$$Q_i = \left\{ y;\ R - a < \left(\sum_{j=1, j\neq i}^{3} y_j^2 \right)^{1/2} < R + a,\ b < |y_i| < \tfrac{1}{2} \right\} \qquad (i = 1, 2, 3),$$

$$Q_1' = Q_1 \backslash \{y;\ y_2 \geqq 0,\ y_3 = 0\}, \qquad Q_2' = Q_2 \backslash \{y;\ y_3 \geqq 0,\ y_1 = 0\},$$

$$Q_3' = Q_3 \backslash \{y;\ y_1 \geqq 0,\ y_2 = 0\},$$

$P_{i,s}$ (resp., $P_{i,f}$) the set of points in P_i such that $R < r < R + a$ (resp., $R - a < r < R$), $Q_{i,s}$ (resp., $Q_{i,f}$) the set of those points in Q_i for which

$$R < \left(\sum_{j=1, j\neq i}^{3} y_j^2 \right)^{1/2} < R + a \left(\text{resp., } R - a < \left(\sum_{j=1, j\neq i}^{3} y_j^2 \right)^{1/2} < R \right).$$

It should be noted that, since a is small enough, we have $Q_{i,s} = Q_i \cap Y_s$ and $Q_{i,f} = Q_i \cap Y_f$.

Finally, set $Q_{i,s}' = Q_i' \cap Q_{i,s}$, $Q_{i,f}' = Q_i' \cap Q_{i,f}$.

We now introduce the functions $G_i : P_i \to Q_i'$ ($i = 1, 2, 3$),

$$G_1(y_1, r, \theta) = (y_1, y_2 = r\cos\theta, y_3 = r\sin\theta),$$

$$G_2(\theta, y_2, r) = (y_1 = r\sin\theta, y_2, y_3 = r\cos\theta),$$

$$G_3(r, \theta, y_3) = (y_1 = r\cos\theta, y_2 = r\sin\theta, y_3).$$

In other words, we associate with each y_i-axis the cylindrical coordinates "around" that axis.

For each i, G_i is a C^1-diffeomorphism of P_i onto Q_i' with Jac $G_i \in (L^\infty(P_i))^9$, Jac $G_i^{-1} \in (L^\infty(Q_i'))^9$ (Jac denotes the Jacobian matrix), $G_i(P_{i,s}) = Q_{i,s}'$, and $G_i(P_{i,f}) = Q_{i,f}'$.

Let us prove the following lemma.

LEMMA A (Extension near the boundary of Y). *For each i there exists an operator $p_i \in \mathscr{L}(H^1(Q_{i,s}), H^1(Q_i))$ such that $p_i u = u$ almost everywhere in $Q_{i,s}$ for all $u \in H^1(Q_{i,s})$.*

Proof. We give the proof, e.g., for $i = 3$. First of all, we define an extension operator in the system of the local coordinates (r, θ, y_3). For $w \in H^1(P_{3,s})$ we put

$$w^*(r, \theta, y_3) = \begin{cases} w(r, \theta, y_3) & \text{if } R < r < R + a, \\ w(2R - r, \theta, y_3) & \text{if } R - a < r < R. \end{cases}$$

Next, for $u \in H^1(Q_{3,s}')$, define $p_3 u = (u \circ G_3)^* \circ G_3^{-1}$, where $(u \circ G_3)(y) = u(G_3(y))$. Clearly, we have just defined a continuous linear operator p_3 of $H^1(Q_{3,s}')$ into $H^1(Q_3')$ such that $p_3 u = u$ almost everywhere in $Q_{3,s}'$ for all $u \in H^1(Q_{3,s}')$. Since the set $\{y; y_1 \geqq 0, y_2 = 0\}$ is of zero measure (with respect to the Lebesgue measure dy), the proof is complete if we verify that p_3 sends $H^1(Q_{3,s})$ into $H^1(Q_3)$. But this follows from the fact that $p_3 u \in H^1(Q_3)$ for all $u \in \mathscr{D}(\bar{Q}_{3,s})$, and use of the density of $\mathscr{D}(\bar{Q}_{3,s})$ in $H^1(Q_{3,s})$ (indeed, the set $Q_{3,s}$ possesses the segment property [1]). □

We are now in a position to construct the very local extension.

The following definition is fundamental to a correct understanding of our purpose.

DEFINITION. A mapping p of $H^1(Y_s)$ into $H^1(Y)$ is said to preserve periodicity if for any u in $H^1(Y_s)$ that takes equal values on opposite faces of Y_s, the function pu also takes equal values on opposite faces of Y (we recall that we are dealing with the case of Fig. 2).

LEMMA B. *There exists an (extension) operator $p \in \mathscr{L}(H^1(Y_s), H^1(Y))$ that preserves periodicity and satisfies the condition $pu = u$ almost everywhere in Y_s for all $u \in H^1(Y_s)$.*

Proof. To begin with, we introduce a C^∞ Y-periodic function ψ on $\mathbb{R}^3$, with $0 \leqq \psi \leqq 1$ and

$$\psi(y) = \begin{cases} 1 & \text{for } y \in \bar{Y}_f, \quad d(y, \Gamma) \leqq a/3, \\ 0 & \text{for } y \in \bar{Y}_f, \quad d(y, \Gamma) \geqq 2a/3 \end{cases}$$

(Γ is defined in § 1; d is the Euclidean metric on $\mathbb{R}^3$).

Next, for $u \in H^1(Y_s)$ we define

$$q(u) = \begin{cases} u & \text{in } Y_s, \\ \psi p_i(u|_{Q_{i,s}}) & \text{in } Y_f \cap \{y; b < |y_i| < \frac{1}{2}\}, \quad i = 1, 2, 3, \end{cases}$$

where actually $p_i(u|_{Q_{i,s}})$ is prolongated by zero in that part of $Y_f \cap \{y; b < |y_i| < \frac{1}{2}\}$ where it is not defined.

Now, we set $Y_f' = Y_f \cap (\cup_{i=1}^3 \{y; b < |y_i| < \frac{1}{2}\})$ and $Y' = Y \setminus (\overline{Y_f \setminus Y_f'})$. Then evidently $q \in \mathscr{L}(H^1(Y_s), H^1(Y'))$ and $q(u) = u$ almost everywhere in Y_s for all $u \in H^1(Y_s)$. Thus the problem reduces to the construction of a continuous linear operator of $H^1(Y')$ into $H^1(Y)$. Observe that the set $Y_f \setminus Y_f'$ is strictly contained in Y. Unfortunately, its interior is not smooth enough and thus we are not in the classical framework of Fig. 1. However, an easy operation can bring us to that classical situation. Indeed, let m_+^i (respectively, m_-^i) be the point in Y_f whose ith coordinate (in the canonical basis of $\mathbb{R}^3$) is b (respectively, $-b$) and the others are zero ($i = 1, 2, 3$). On the other hand, let $B = B(0, R)$ denote the open ball in $\mathbb{R}^3$ centred at the origin, with radius R. Then set $B_+^i = \{y \in B + m_+^i, y_i > b\}$, $B_-^i = \{y \in B + m_-^i, y_i < -b\}$. Since R is small enough compared to $\frac{1}{2}$ (see § 1) and moreover the number b can be chosen so that $b + R < \frac{1}{2}$, the

above sets are contained in Y_f. Moreover, they are parts of Y_f'. Finally, define $D = (Y_f \setminus Y_f') \cup [\cup_{i=1}^{3} (B_+^i \cup B_-^i)]$, which is an open set of class C^1, with closure contained in Y, and $(Y \setminus \bar{D}) \subset Y'$. Then it is an easy task (see, e.g., [1]) to construct $p' \in \mathcal{L}(H^1(Y \setminus \bar{D}), H^1(Y))$ such that $p'u = u$ almost everywhere in $Y \setminus \bar{D}$ for all $u \in H^1(Y \setminus \bar{D})$. Denoting by $\mathcal{R}$ the restriction operator $w \to w|_{(Y \setminus \bar{D})}$, we finally define the desired operator p as

$$pu = p'(\mathcal{R}(qu)) \quad \text{for } u \in H^1(Y_s).$$

The proof is complete. □

Remark. We need not insist upon the fact that the operator p preserves periodicity. The nice operators p_i $(i = 1, 2, 3)$ in Lemma A (see, in particular, the mapping $w \to w^*$ and its analogues relative to $i = 1, 2$) and q in Lemma B (see the appropriate function ψ) were specially constructed for that purpose.

Acknowledgment. The author acknowledges the referee for kind advice and for some helpful remarks.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] A. BENSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.
[3] D. CIORANESCU AND F. MURAT, *Un terme étrange venu d'ailleurs*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vols. II, III, Pitman, Boston, 1980.
[4] D. CIORANESCU AND J. SAINT-JEAN PAULIN, *Homogenization in open sets with holes*, J. Math. Anal. Appl., 71 (1979), pp. 590-607.
[5] C. CONCA, *On the application of the homogenization theory to a class of problems arising in fluid mechanics*, Publ. Lab. Anal. Numer., 83006 (1983), pp. 1-63.
[6] H. I. ENE AND E. SANCHEZ-PALENCIA, *Equations et phénomènes de surface pour l'écoulement dans un milieu poreux*, J. Méc. Théor. Appl., 14 (1975), pp. 73-108.
[7] F. FLEURY, *Propagation des ondes dans une suspension de particules solides*, C. R. Acad. Sci. Paris Sér. A, 288 (1979), pp. 77-80.
[8] H. G. GARNIR, *Les problèmes aux limites de la physique mathématique*, Birkhäuser, Basel, 1958.
[9] E. HILLE AND R. S. PHILLIPS, *Functional analysis and semigroups*, Amer. Math. Soc. Colloq. Publ., 31 (1957), pp. 19-664.
[10] I. HLAVAČEK AND J. NEČAS, *On inequalities of Korn's type and application to linear elasticity*, Arch. Rational Mech. Anal., 36 (1970), pp. 312-334.
[11] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.
[12] L. LANDAU AND E. LIFCHITZ, *Mécanique des Fluides*, Mir, Moscow, 1971.
[13] T. LEVY, *Propagation of waves in a fluid-saturated porous elastic solid*, Internat. J. Engrg. Sci., 17 (1979), pp. 1005-1014.
[14] J. L. LIONS, *Perturbations singulières dans les problèmes aux limites et en contrôle optimal*, Lecture Notes in Math. 323, Springer-Verlag, Berlin, New York, 1973.
[15] G. NGUETSENG, *Etude asymptotique du comportement macroscopique d'un mélange de deux fluides visqueux*, J. Méc. Théor. Appl., 1 (1982), pp. 951-961.
[16] ———, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608-623.
[17] G. NGUETSENG AND E. SANCHEZ-PALENCIA, *On the asymptotics of the vibration problem for a solid-fluid mixture*, Bull. Sci. Math. (2), 107 (1983), pp. 413-435.
[18] E. SANCHEZ-PALENCIA, *Non-Homogeneous Media and Vibration Theory*, Lecture Notes in Physics 127, Springer-Verlag, Berlin, New York, 1980.
[19] J. SANCHEZ-HUBERT AND E. SANCHEZ-PALENCIA, *Existence et unicité de la solution d'un problème de vibration d'un mélange de solide et fluide*, C. R. Acad. Sci. Paris Sér. I, 294 (1982), pp. 107-109.
[20] J. SANCHEZ-HUBERT, *Asymptotic study of the macroscopic behaviour of a solid-fluid mixture*, Math. Meth. Appl. Sci., 2 (1980), pp. 1-11.

[21] L. SCHWARTZ, *Méthodes Mathématiques pour les Sciences Physiques*, Hermann, Paris, 1965.
[22] L. TARTAR, *Homogénéisation en hydrodynamique*, in Singular Perturbations and Boundary Layer Theory, Lecture Notes in Math. 594, Springer-Verlag, Berlin, New York, 1977.
[23] ———, *Problèmes d'homogénéisation dans les équations aux dérivées partielles*, Cours peccot, Collège de France, Paris, 1977.
[24] R. TEMAM, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1977.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1415-1425, November 1990
© 1990 Society for Industrial and Applied Mathematics
004

FINITE-TIME BLOWUP FOR A PARTICULAR PARABOLIC SYSTEM*

J. BEBERNES† AND A. LACEY‡

Abstract. The problem of blowup in finite time is considered for an initial boundary value problem for a two-dimensional parabolic system. This system models exothermic chemical reactions taking place within a porous medium, assuming one diffusing reactant and the usual Frank-Kamenetskii approximation to the classical Arrhenius rate law.

Key words. blowup, parabolic systems, self-similar solutions

AMS(MOS) subject classifications. 35J, 35K, 35K55, 35J60

1. Introduction. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with smooth boundary $\partial\Omega$. For the initial boundary value problem

$$\text{(S)}\qquad \begin{aligned} &u_t - \Delta u = f(x, t, u), && x \in \Omega, \quad t > 0, \\ &u(x, 0) = u_0(x), && x \in \Omega, \\ &\frac{\partial u}{\partial n} + \mu u(x, t) = 0, && x \in \partial\Omega, \quad t > 0, \end{aligned}$$

it is well known that classical solutions may cease to exist by becoming unbounded in some norm as a finite maximal time T is approached [1], [5], [8]. This phenomenon is called blowup. For a large class of nonlinearities, an active area of research has been to determine and characterize how solutions of (S) blow up.

For parabolic systems of the form

$$\text{(C)}\qquad \begin{aligned} &u_t - \Delta u = f(u, v), \\ &pv_t - \Delta v = g(u, v), \end{aligned} \qquad x \in \Omega, \quad t > 0,$$

$$u(x, 0) = u_0(x), \quad v(x, 0) = v_0(x), \quad x \in \Omega,$$

$$\frac{\partial u}{\partial n} + \mu u = 0, \quad \frac{\partial v}{\partial n} + \nu v = 0, \quad x \in \partial\Omega, \quad t > 0,$$

where $p > 0$, $\mu, \nu \in [0, \infty]$, considerably less is known as to when solutions exist globally or cease to exist in finite time [4], [7], [10].

If $f(u, v) = u^\beta v$ and $g(u, v) = -u^\beta v$ with $\beta \geqq 1$, then Hollis, Martin, and Pierre [7] have shown that solutions of (C) exist globally. If $f(u) = A e^{\alpha u}$ and $g(v) = B e^{\beta v}$, A, B, α, β positive, then Friedman and Giga [6] have extended the known results [5] for a single scalar equation to such systems to show single-point finite-time blowup in the one-dimensional symmetric spatial case with $\mu = \nu = \infty$.

* Received by the editors March 27, 1989; accepted for publication (in revised form) November 22, 1989. This research was partially supported by U.S. Army Research Office contract DAAL 03-88-K0111 and by U.K. Science and Engineering Council grant GR/D/73096.

† Department of Mathematics, University of Colorado, Boulder, Colorado 80309.

‡ Department of Mathematics, Heriot-Watt University, Riccarton, Edinburgh EH14 4AS, United Kingdom.

The purpose of this paper is to investigate the blowup in finite time of solutions of special forms of the initial boundary value problem:

$$
\begin{aligned}
&u_t - \Delta u = \lambda(1-v)f(u),\\
&pv_t - \Delta v = \alpha\lambda(1-v)f(u),\\
&u(x,0) = u_0(x), \qquad v(x,0) = v_0(x),\\
&\frac{\partial u}{\partial n} + \mu u = 0, \qquad \frac{\partial v}{\partial n} + \nu v = 0.
\end{aligned} \tag{E}
$$

Problem (E) arises as an approximating model for an exothermic chemical reaction taking place within a porous medium, assuming one diffusing reactant and the usual Frank-Kamenetskii approximation $f(u) = e^u$ to the classical Arrhenius rate law. The variables u and v are chosen so that physically the scaled temperature of the reaction process above ambient is u and the scaled concentration is $1-v$.

In [4], Burnell, Lacey, and Wake have investigated the steady states of system (E) with $f(u) = e^u$. If $\mu = \nu = \infty$ or if $\nu < \infty$, $\mu > 0$, then steady-state solutions exist for all $\lambda > 0$. If $\nu = \infty$, $0 < \mu < \infty$, then for spatial dimensions $n = 1$, 2, or 3 there exists a critical value $\lambda^* > 0$ such that no steady-state solution exists for $\lambda > \lambda^*$ and steady-state solutions do exist for $0 < \lambda < \lambda^*$. For $\nu = \infty$ and $\mu = 0$, no steady-state solution exists for any $\lambda > 0 \equiv \lambda^*$.

For $\lambda > \lambda^*$, we might conjecture that solutions to (E) do not exist globally and hence blow up in finite time in some sense. In fact, it seems plausible that for $\lambda > \lambda^*$, the heat produced by the incoming reactant cannot be diffused through the boundary sufficiently quickly so that the solution to (E) becomes unbounded in finite or infinite time.

The purpose of this paper is to prove that the solution $(u(x,t), v(x,t))$ of (E) blows up in finite time T in the very special case of one spatial variable $n = 1$ with $\mu = 0$, $\nu = \infty$ (so Neumann boundary condition for u and Dirichlet for v), $f(u) = e^u$, $p = 1$, and $\lambda^* = 0$.

In a forthcoming paper, we will consider generalizations to higher spatial dimensions $n > 1$, $\nu = \infty$, $0 < \mu < \infty$, $f(u) \not\equiv e^u$, and $\lambda > \lambda^* > 0$.

2. Preliminaries and statement of theorem. Consider the parabolic system

$$
u_t - u_{xx} = (1-v)\,e^u, \qquad v_t - v_{xx} = (1-v)\,e^u \tag{1}
$$

on $\Omega \times [0, \infty)$, where $\Omega = (0, 1)$ with the initial boundary conditions

$$
\begin{aligned}
&u(x,0) = u_0(x), \qquad v(x,0) = v_0(x),\\
&\frac{\partial u}{\partial n}(x,t) = 0, \quad v(x,t) = 0, \quad x \in \partial\Omega, \quad t > 0
\end{aligned} \tag{2}
$$

with

$$
\Delta u_0(x) + \lambda(1 - v_0(x))\,e^{u_0(x)} \geqq 0, \quad u_0(x) \geqq 0, \quad v_0(x) \geqq 0, \quad \Delta u_0 - \Delta v_0 \geqq 0 \tag{3}
$$

where $u_0(x)$, $v_0(x)$ are symmetric on Ω.

Although we are taking $\alpha = \lambda = 1$, other positive values can be taken in a similar way. We use (3) to give monotone solutions, but we can use these to show blowup for more general initial data by comparison with smaller monotone increasing solutions (in the u, h formulation given below).

By setting $h = u - v$, we have

$$u_t - u_{xx} = (1 + h - u)\, e^u, \quad h_t - h_{xx} = 0 \tag{4}$$

with

$$\begin{aligned} &u(x, 0) = u_0(x), \qquad h(x, 0) = u_0(x) - v_0(x), \\ &\frac{\partial u}{\partial n}(x, t) = 0, \quad h(x, t) = u(x, t), \quad x \in \partial\Omega, \quad t > 0. \end{aligned} \tag{5}$$

Set

$$\begin{aligned} &Hw = w_t - w_{xx}, \\ &\pi_\sigma = \{(x, t): 0 < x < 1,\ 0 < t < \sigma\}. \end{aligned}$$

The following facts [1], [6] are known:

1. There exists a unique classical solution (u, v) of (1)-(2) on π_σ for some $\sigma > 0$. Equivalently, there exists a unique classical solution (u, h) of (4)-(5) on π_σ, $\sigma > 0$. Let $T \equiv \sup\{\sigma: (u, v) \text{ exists on } \pi_\sigma\}$.

2. $u(x, t) \geqq 0$, $v(x, t) \geqq 0$ on π_T.

3. $u_t(x, t) > 0$ on π_ε for some $\varepsilon > 0$.

By property 3 and the maximum principle, $h_t > 0$ on π_ε. Then noting that (u_t, h_t) satisfies

$$\begin{aligned} &h_{tt} = h_{xxt}, \qquad u_{tt} - u_{txx} + (u - h)\, e^u u_t = e^u h_t, \\ &u_t(x, 0) \geqq 0, \qquad h_t(x, 0) \geqq 0, \\ &\frac{\partial u_t}{\partial n}(x, t) = 0, \qquad h_t(x, t) = u_t(x, t) \geqq 0 \quad \text{on } \partial\Omega \times (0, \varepsilon), \end{aligned} \tag{6}$$

we can again apply the maximum principle to u_t and h_t to assert $u_t \geqq 0$ and $h_t \geqq 0$ on $\bar{\pi}_\varepsilon$ and hence on π_T.

4. $u_t(x, t) \geqq 0$, $h_t(x, t) \geqq 0$ on π_T.

5. If $T = +\infty$, then $u(x, t) \nearrow +\infty$, $h(x, t) \nearrow +\infty$ for all $x \in \Omega$ as $t \to +\infty$.

In the following sections, we will prove a sequence of results that lead to the principal theorem of this paper.

THEOREM 2.1. *The solution $(u(x, t), v(x, t))$ of the initial boundary value problem (1), (2) blows up in finite time $T < \infty$ in the sense that*

$$\lim_{t \to T^-} \left[\sup_\Omega u(x, t) + \sup_\Omega v(x, t) \right] = +\infty.$$

3. Idea of the proof and shifting the problem. Assume that $T = +\infty$. This means the solution (u, h) of the initial boundary value problem (4), (5) exists globally and we have no finite-time blowup. From property 5 above we have that

$$u(x, t) \to \infty, \quad h(x, t) \to \infty \quad \text{as } t \to +\infty \quad \text{for all } x \in \Omega.$$

For each natural number N, define

$$T_N \equiv \inf\{t > 0: u(x, s) > N,\ h(x, s) > N \text{ for all } x \in \Omega,\ s > t\}.$$

Then $\{T_N\}_1^\infty$ is a sequence of positive real numbers such that

$$\begin{aligned} &0 < T_N < T_{N+1}, \\ &u(x, T_N) \geqq N, \qquad h(x, T_N) \geqq N \quad \text{for all } x \in \Omega. \end{aligned}$$

Then

$$T_1 + \sum_{N=1}^{\infty} (T_{N+1} - T_N) = T = +\infty$$

by our assumption and hence $\sum_{N=1}^{\infty} (T_{N+1} - T_N)$ diverges.

Our goal in this and the following sections is to shift the problem in such a way as to allow us to construct a sequence $\{\tau_N\}$ with the property that

$$\sum_N \tau_N < \infty, \qquad T_{N+1} - T_N \leqq \tau_N.$$

This clearly contradicts $T = +\infty$. We conclude finite-time blowup.

We now begin to shift the problem by translating the dependent variables by N and then translating the time by T_N.

For each positive integer N, let (u_N, h_N) be the unique solution of

$$(7_N) \qquad u_t - u_{xx} = (1 + h - u)\, e^{u+N}, \qquad h_t - h_{xx} = 0$$

with

$$(8_N) \qquad \begin{aligned} &u(x, 0) = 0 = h(x, 0), \qquad x \in \Omega, \\ &\frac{\partial u}{\partial n}(x, t) = 0, \quad u(x, t) = h(x, t), \quad x \in \partial\Omega, \quad t > 0. \end{aligned}$$

The solution (u_N, h_N) satisfies properties 1-5 with $\hat{T}_N$ denoting the right-hand endpoint of maximal interval of existence.

THEOREM 3.1. *For $N \in \mathbb{N}$, let (u_N, h_N) be the solution of (7_N)-(8_N). Let (u, h) be the solution of* (4), (5). *Then*

$$u_N(x, t) \leqq u_N(x, t) + N \leqq u(x, t + T_N),$$
$$h_N(x, t) \leqq h_N(x, t) + N \leqq h(x, t + T_N)$$

on π_∞.

Proof. Let $(\alpha_1, \alpha_2) = (u_N + N, h_N + N)$, where (u_N, h_N) is the solution of (7_N), (8_N). Then (α_1, α_2) solves (4) with initial boundary conditions

$$(9) \qquad \begin{aligned} &\alpha_1(x, t) = \alpha_2(x, t), \quad \frac{\partial \alpha_1}{\partial n} = 0, \quad x \in \partial\Omega, \\ &\alpha_1(x, 0) = \alpha_2(x, 0) = N, \qquad x \in \Omega. \end{aligned}$$

Since $(u(x, t + T_N), h(x, t + T_N))$ solves (4) with initial boundary conditions

$$(10) \qquad \begin{aligned} &u(x, t + T_N) = h(x, t + T_N), \quad \frac{\partial u(x, t + T_N)}{\partial n} = 0, \quad x \in \partial\Omega, \\ &u(x, T_N) \geqq N, \quad h(x, T_N) \geqq N, \quad x \in \Omega, \end{aligned}$$

$(u(x, t + T_N), h(x, t + T_N))$ is an upper solution for (4) and (9). Thus,

$$u_N(x, t) \leqq u_N(x, t) + N \leqq u(x, t + T_N),$$
$$h_N(x, t) \leqq h_N(x, t) + N \leqq h(x, t + T_N)$$

on π_∞.

COROLLARY. *$\hat{T}_N = +\infty$ and $u_N(x, t) \to \infty$, $h_N(x, t) \to \infty$ as $t \to +\infty$.*

This follows from the definition of u_N and h_N and a comparison with u and h. Let $t_N \equiv \inf\{t > 0: h_N(x, s) > 1 \text{ for all } x \in \Omega, s > t\}$; then

$$h_N(x, t_N) \geqq 1 \quad \text{for all } x \in \Omega,$$
$$T_{N+1} \leqq T_N + t_N.$$

We will now begin to get upper-bound estimates on t_N.

4. Construction of lower solutions. As before, let $\Omega = (0, 1)$ and let $\Gamma = (0, \infty)$ be a half line in $\mathbb{R}$ containing Ω.

THEOREM 4.1. *For a given $N \in \mathbb{N}$, let (u, h) be the solution of (7_N), (8_N) on π_∞ and let $(\tilde{u}, \tilde{h})$ be the solution of (7_N), (8_N) on $\Gamma \times (0, \infty)$; then*

$$\begin{aligned} \tilde{u}(x, t) &\leqq u(x, t) \\ \tilde{h}(x, t) &\leqq h(x, t) \end{aligned} \quad \text{for } (x, t) \in \pi_\infty.$$

Proof. We proceed by proving a sequence of lemmas.

LEMMA 1. *Given $f(t)$, f increasing on $[0, \infty)$ with $f(0) = 0$, let $h(x, t)$ and $\tilde{h}(x, t)$ be solutions of*

$$\begin{aligned} &z_t = z_{xx}, \quad z(x, 0) = 0, \quad z(x, t) = f(t), \\ &x \in \partial\Omega \qquad (x \in \partial\Gamma) \end{aligned} \tag{11}$$

on π_∞ and $\Gamma \times (0, \infty)$, respectively. Then

$$h(x, t) \geqq \tilde{h}(x, t) \quad \text{on } \pi_\infty.$$

LEMMA 2. *Let $h_0(x, t) \equiv 0$, $u_0(x, t) \equiv 0$; then the solutions $u_1(x, t)$ and $\tilde{u}_1(x, t)$ of*

$$u_t - u_{xx} = e^N, \quad \frac{\partial u}{\partial n} = 0, \quad u(x, 0) = 0 \tag{12}$$

on π_∞ and $\Gamma \times (0, \infty)$, respectively, satisfy

$$u_1(x, t) \geqq \tilde{u}_1(x, t) \quad \text{on } \pi_\infty.$$

For

$$h_t - h_{xx} = 0, \quad h(x, 0) = 0, \quad h(x, t) = u_1(x, t) \tag{13_1}$$

let $h_1(u_1)$ and $\tilde{h}_1(u_1)$ denote the solutions of (13_1) on π_∞ and $\Gamma \times (0, \infty)$, respectively. Then, by Lemma 1, using the fact that $u_1(0, t) = u_1(+1, t) = f(t)$ by symmetry,

$$h_1(u_1) \geqq \tilde{h}_1(u_1) \quad \text{on } \pi_\infty,$$

and by a standard comparison theorem

$$\tilde{h}_1(u_1) \geqq \tilde{h}_1(\tilde{u}_1) \quad \text{on } \pi_\infty.$$

By the maximum principle,

$$h_1 - \tilde{h}_1 \geqq u_1 - \tilde{u}_1 \quad \text{on } \pi_\infty. \tag{14_1}$$

Next consider

$$\begin{aligned} &u_t - u_{xx} = e^N (1 + h_1(u_1) - u_1)\, e^{u_1} \quad \text{on } \pi_\infty, \\ &u(x, 0) = 0, \qquad \frac{\partial u}{\partial n}(x, t) = 0, \end{aligned} \tag{15_1}$$

$$u_t - u_{xx} = e^N (1 + \tilde{h}_1(\tilde{u}_1) - \tilde{u}_1)\, e^{\tilde{u}_1}, \qquad \Gamma \times (0, \infty),$$

$$u(x, 0) = 0, \qquad \frac{\partial u}{\partial n}(x, t) = 0. \tag{16_1}$$

Let $u_2(x, t)$ be the solution of the standard modification of (15_1) needed for the monotone method, and let $\tilde{u}_2(x, t)$ be the solution of a similar modification of (16_1) on $\Gamma \times (0, \infty)$ (see Pao [10] or [11]). Then $u_2(x, t) \geqq u_1(x, t)$ on π_∞ and $\tilde{u}_2(x, t) \geqq \tilde{u}_1(x, t)$ on $\Gamma \times (0, \infty)$. By (14_1)

$$u_2(x, t) \geqq \tilde{u}_2(x, t) \quad \text{on } \pi_\infty.$$

We now proceed inductively, considering (13_k), (14_k), (15_k), and (16_k) to get Lemma 3.

LEMMA 3. (a) $h_k(u_k) \geqq \tilde{h}_k(\tilde{u}_k)$ *on* π_∞ *with*

$$h_k(u_k) \geqq h_{k-1}(u_{k-1}) \quad \text{on } \pi_\infty \quad \text{and} \quad \tilde{h}_k(\tilde{u}_k) \geqq \tilde{h}_{k-1}(\tilde{u}_{k-1}) \quad \text{on } \Gamma \times (0, \infty).$$

(b) $u_k \geqq \tilde{u}_k$ *on* π_∞ *with*

$$u_k \geqq u_{k-1} \quad \text{on } \pi_\infty \quad \text{and} \quad \tilde{u}_k \geqq \tilde{u}_{k-1} \quad \text{on } \Gamma \times (0, \infty).$$

We now use a standard monotone argument to prove that $u_k \to u$, $\tilde{u}_k \to \tilde{u}$, $h_k \to h$, $\tilde{h}_k \to \tilde{h}$, where (u, h) and $(\tilde{u}, \tilde{h})$ are solutions of (7_N)-(8_N) on π_∞ and $\Gamma \times (0, \infty)$, respectively, with

$$u(x, t) \geqq \tilde{u}(x, t), \quad h(x, t) \geqq \tilde{h}(x, t) \quad \text{on } \pi_\infty.$$

This completes the proof of Theorem 4.1.

For a given $N \in \mathbb{N}$, let $(\tilde{u}(x, t), \tilde{h}(x, t))$ be the solution of (7_N), (8_N) on $\Gamma \times (0, \infty)$. Set

$$x = e^{-N/2} y \qquad t = e^{-N} \tau;$$

then

$$\tilde{U}(y, \tau) = \tilde{u}(x(y), t(\tau)), \qquad \tilde{H}(y, \tau) = \tilde{h}(x(y), t(\tau));$$

solve

$$(17_N) \qquad \tilde{U}_\tau - \tilde{U}_{yy} = (1 + \tilde{H} - \tilde{U})\, e^{\tilde{U}}, \qquad \tilde{H}_\tau - \tilde{H}_{yy} = 0,$$

$$\tilde{H}(y, 0) = \tilde{U}(y, 0) = 0,$$

$$(18_N) \qquad \tilde{H}(y, \tau) = \tilde{U}(y, \tau), \quad \frac{\partial \tilde{U}}{\partial n} = 0, \quad y \in \partial\Gamma, \quad \tau \geqq 0.$$

Set $f(\tau) = \tilde{H}(0, \tau) = \tilde{U}(0, \tau)$; then

$$f(e^N t) = \tilde{u}(0, t) = \tilde{h}(0, t),$$

$$(19_N) \qquad \tilde{h}(x, t) = \int_0^t -2k_x(x, t - \tau) f(e^N \tau)\, d\tau$$

where

$$k(x, t) = \frac{1}{\sqrt{4\pi t}}\, e^{-x^2/4t}$$

is the heat kernel. We now proceed to obtain reasonably sharp lower bounds on $f(\tau)$, which we begin to do by constructing lower solutions for (17), (18) as follows. Consider

$$(20) \qquad w^0_\tau - w^0_{yy} = (1 - w^0), \qquad z^0_\tau - z^0_{yy} = 0$$

with

$$w^0(y,0)=z^0(y,0)=0,$$
$$w^0(0,\tau)=0, \qquad \frac{\partial(w^0+z^0)}{\partial n}(0,\tau)=0. \tag{21}$$

We first observe that this system is essentially decoupled. This allows us to prove the following theorem.

THEOREM 4.2. *The solution* $w^0(y,\tau)$ *of*

$$w^0_\tau - w^0_{yy} = 1-w^0,$$
$$w^0(y,0)=0, \qquad w^0(0,\tau)=0 \tag{22}$$

satisfies:

(a) $0 \leqq w^0(y,\tau) \leqq 1-e^{-y}$ *on* $(0,\infty)\times(0,\infty)$,

(b) $w^0_\tau(y,\tau)\geqq 0$, $-(\partial w^0/\partial n)(0,\tau)\leqq 1$,

(c) $w^0(y,\tau)\nearrow 1-e^{-y}$ *as* $\tau\to\infty$.

Proof. We simply note that $\alpha(y,\tau)\equiv 0$ is a lower solution for (22) and that the steady-state solution $\beta(y)=1-e^{-y}$ of

$$-u_{yy}=1-u, \qquad u(0)=0 \tag{23}$$

is an upper solution for (22). Hence (a) follows.

Conclusion (b) follows by observing that $v(y,\tau)\equiv w^0(y,\tau+h)$ is an upper solution of (22) on π_∞. Thus $(w^0(y,\tau+h)-w^0(y,\tau))/h>0$, from which we have $w^0_\tau(y,\tau)\geqq 0$.

Conclusion (c) follows by standard arguments (see, for example, [3] or [8]).

COROLLARY. *For any* $c\in(0,1)$, *there exists* $\tau_c>0$ *such that* $-(\partial w^0/\partial n)(0,\tau)>c$ *for* $\tau>\tau_c$ *and hence* $(\partial z^0/\partial n)(0,\tau)>c$ *for* $\tau>\tau_c$.

For $\tau>\tau_c$, let $s(y,\tau)=c(\tau-\tau_c)^{1/2}g(y/(\tau-\tau_c)^{1/2})$ be the self-similar solution of $z_\tau - z_{yy}=0$ with $(\partial s/\partial n)(0,\tau)=c$, $s(y,0)=0$.

THEOREM 4.3. *For any* $L<\infty$, $c_2>0$ *with* $c_2<cg(0)=2c/\sqrt{\pi}$,

$$z^0(y,\tau)\geqq c_2\tau^{1/2} \quad \text{as } \tau\to\infty \quad \text{if } 0\leqq y\leqq L. \tag{24}$$

Proof. From the maximum principle, $z^0(y,\tau)\geqq s(y,\tau)$. Taking $0<c_2<cg(0)$, where $g(0)=2/\sqrt{\pi}$, then for any fixed y,

$$s(y,\tau)\sim c(\tau-\tau_c)^{1/2}g(0)\sim cg(0)\tau^{1/2}>c_2\tau^{1/2}$$

for $\tau\to\infty$. Thus, for any given $L>0$, $z^0(y,\tau)\geqq s(L,\tau)>c_2(\tau^{1/2})$ as $\tau\to\infty$ for $0\leqq y\leqq L$.

THEOREM 4.4. *The solution* (w^0,z^0) *of* (20), (21) *has the property that* (w^0+z^0,z^0) *is a lower solution for* (17), (18). *Thus* $\tilde{U}\geqq w^0+z^0$ *and* $\tilde{H}\geqq z^0$ *on* $\Gamma\times(0,\infty)$.

Proof. The result is immediate since (w^0+z^0,z^0) satisfies

$$z^0_\tau+z^0_{yy}=0,$$
$$(z^0+w^0)_\tau+(z^0+w^0)_{yy}=1-w^0\leqq(1+z^0-(w^0+z^0))\,e^{w^0+z^0}$$

with $(\partial/\partial n)(z^0+w^0)(0,\tau)=0$, $(z^0+w^0)(y,0)=0$, $z^0(y,0)=0$.

By Theorem 4.4, we have that

$$f(\tau)=\tilde{H}(0,\tau)\geqq z^0(0,\tau)$$

and by (24)

$$z^0(0,\tau)\geqq c_2\tau^{1/2}$$

for τ sufficiently large.

We have, using (19_N),

$$\tilde{h}(x, t) \geqq \hat{h}(x, t) = -2c_2 \int_0^t k_x(x, t-s)\, e^{N/2} s^{1/2}\, ds$$

where $\hat{h}(x, t)$ is the solution to $\hat{h}_t = \hat{h}_{xx}$, $\hat{h}(0, t) = c_2\, e^{N/2} t^{1/2}$, $\hat{h}(x, 0) = 0$, and is monotone decreasing in x. Then for $0 \leqq x \leqq 1$,

$$\begin{aligned}\hat{h}(x, t) \geqq \hat{h}(1, t) &= \frac{c_2\, e^{N/2}}{2\sqrt{\pi}} \int_0^t (t-s)^{-3/2} \exp\left[-\frac{1}{4(t-s)}\right] s^{1/2}\, ds \\ &\geqq \frac{c_2}{2\sqrt{\pi}} \int_0^1 s^{1/2}(1-s)^{-3/2} \exp\left[N\left(\frac{1}{2} - (4K(1-s))^{-1}\right)\right] ds, \\ &\geqq 1 \qquad \text{taking } t = K/N\end{aligned}$$

for $K > \frac{1}{2}$ and N large since the dominant contribution of the integral for $N \gg 1$ comes near $s = 0$, where $\frac{1}{2} - 1/4K(1-s) = (K - \frac{1}{2})/2K$. It follows that $\tilde{h}(x, t) \geqq 1$ for $t \geqq t_N$, where $t_N \leqq \tau_N = K/N$ for some $K \geqq \frac{1}{2}$.

The upper bound K/N is not sufficiently sharp to give finite-time blowup. We can, however, improve the estimate by finding sharper lower solutions. To do this, we consider

$$\begin{aligned} &w'_\tau - w'_{yy} = (1 - w')\, e^{z_0(y,\tau)} \\ &z'_\tau - z'_{yy} = 0 \end{aligned} \qquad (0, L) \times (0, \infty) \tag{25}$$

with

$$\begin{aligned} &w'(0, \tau) = 0, \qquad w'(L, \tau) = w^0(L, \tau), \\ &w'(y, 0) = 0, \qquad y \in (0, L), \\ &z'(y, 0) = 0, \qquad y > 0, \\ &-\frac{\partial z'}{\partial n} = \frac{\partial w'}{\partial n}(0, \tau), \qquad \tau > 0. \end{aligned} \tag{26}$$

THEOREM 4.5. *The solution* (w', z') *of* (25), (26) *has the property that* $(w' + z', z')$ *is a lower solution for* (17), (18) *on* $(0, L) \times (0, \infty)$.

Proof. We simply observe that $(w' + z', z')$ satisfies

$$\begin{aligned} &z'_\tau - z'_{yy} = 0, \\ &(z' + w')_\tau - (z' + w')_{yy} = w'_\tau - w'_{yy} = (1 - w')\, e^{z^0(y,\tau)} \\ &\qquad\qquad \leqq (1 + z' - (w' + z'))\, e^{z' + w'} \end{aligned}$$

since $w'(y, \tau) \geqq w^0(y, \tau) \geqq 0$ and hence $z'(y, \tau) \geqq z^0(y, \tau)$ on $(0, L) \times (0, \infty)$. Clearly, $z'(y, 0) = (z' + w')(y, 0) = 0$ for $y \in (0, L)$, $z'(0, \tau) = z'(0, \tau) + w'(0, \tau)$, and $(\partial/\partial n)(z' + w')(0, \tau) = 0$. Thus $\tilde{H}(y, \tau) \geqq z'(y, \tau)$ and $\tilde{U}(y, \tau) \geqq (z' + w')(y, \tau)$ on $(0, L) \times (0, \infty)$. The conclusion is immediate.

5. Completion of the proof. Using the lower solution $(z' + w', z')$ of (17), (18) just constructed, we can now get an improved lower bound on $z'(y, \tau)$.

THEOREM 5.1. *For* $y \in (0, \bar{L})$, $0 < \bar{L} < L$, *and* τ *sufficiently large, the solution* (w', z') *of* (25), (26) *satisfies*

$$z'(y, \tau) \geqq c_3 \tau^{1/2}\, e^{c_4 \tau^{1/2}}. \tag{27}$$

Proof. By (24), there exists $K>0$ such that $z^0(y,\tau)\geqq c_2\tau^{1/2}$ for $0\leqq y\leqq L$ and $\tau\geqq K$. This in turn implies that the solution $(w'(y,\tau), z'(y,\tau))$ of (25), (26) has the property that $w'(y,\tau)$ satisfies

$$\begin{aligned} w'_\tau - w'_{yy} &\geqq e^{c_2\tau^{1/2}}(1-w') \\ &\geqq e^{c_2 J^{1/2}}(1-w') \quad \text{for } \tau\geqq J\geqq K \end{aligned}$$

with $w'(0,\tau)=0$, $w'(L,\tau)=w^0(L,\tau)$, and $w'(y,0)=0$.

Consider

$$\begin{aligned} &\theta_\tau - \theta_{yy} = e^{c_2 J^{1/2}}(1-\theta), \\ &\theta(y,J)=0, \qquad \theta(0,\tau)=0 \end{aligned} \tag{28}$$

and rescale, letting

$$s = e^{c_2 J^{1/2}}(\tau - J), \qquad z = e^{c_2 J^{1/2}/2} y;$$

then

$$\bar\theta(z,s) = \theta(y,\tau)$$

satisfies

$$\begin{aligned} &\bar\theta_s - \bar\theta_{zz} = (1-\bar\theta), \\ &\bar\theta(z,0)=0, \qquad \bar\theta(0,s)=0. \end{aligned} \tag{29}$$

As in Theorem 4.3, given $c_3\in(0,1)$, there exists $\tau_{c_3}>0$ such that

$$\bar\theta_z(0,s) > c_3 \quad \text{for } s\geqq \tau_{c_3},$$

and hence

$$\theta_y(0,\tau)\geqq c_3\, e^{c_2 J^{1/2}/2} \quad \text{for } \tau\geqq \tau_{c_3}\, e^{-c_2 J^{1/2}} + J.$$

Since $w'(y,\tau)$ is an upper solution for (28),

$$w'_y(0,\tau)\geqq c_3\, e^{c_2 J^{1/2}/2} \quad \text{for } \tau\geqq \tau_{c_3}\, e^{-c_2 J^{1/2}} + J.$$

Using the same idea as in the proof of Theorem 4.3, from

$$-z'_y(y,\tau)\geqq c_3\, e^{c_2 J^{1/2}/2} \quad \text{for } \tau\geqq J + \tau_{c_3}\, e^{c_2 J^{1/2}}$$

we have

$$z'(y,\tau)\geqq K' c_3\, e^{c_2 J^{1/2}/2}(\tau - J)^{1/2} \quad \text{for } 0\leqq y\leqq L.$$

In particular, taking $\tau = 2J$,

$$z'(y,\tau)\geqq K' c_3 J^{1/2}\, e^{(c_2/2)J^{1/2}} = (2^{-1/2} c_3 K')\tau^{1/2}\, e^{(2^{-3/2}c_2)\tau^{1/2}}$$

for $\tau/2 = J \geqq K' > K$, where K' is chosen to ensure $J\geqq \tau_{c_3}\, e^{-c_2 J^{1/2}}$. Thus,

$$z'(y,\tau)\geqq c_3\tau^{1/2}\, e^{c_4\tau^{1/2}}, \tag{27}$$

redefining c_3.

This lower-bound estimate (27) for $z'(y,\tau)$ can now be used to get an improved upper bound for t_N.

By Theorem 4.5,

$$f(\tau) = \tilde H(0,\tau)\geqq z'(0,\tau),$$

and by (27),

$$z'(0,\tau) \geqq c_3 \tau^{1/2} e^{c_4 \tau^{1/2}}, \qquad \tau \geqq 2K'.$$

Thus $\tilde{H}(0,\tau) = f(\tau) \geqq z'(0,\tau) \geqq c_3 \tau^{1/2} e^{c_4 \tau^{1/2}}$ and hence

$$\tilde{h}(0,t) = f(e^N t) \geqq c_3 e^{N/2} t^{1/2} e^{c_4 e^{N/2} t^{1/2}}, \qquad t \geqq 2K' e^{-N}. \tag{30}$$

The solution $\tilde{h}(x,t)$ of the problem

$$h_t - h_{xx} = 0, \quad h(0,t) = f(e^N t), \quad h(x,0) = 0 \quad \text{for } x > 0 \tag{31}$$

can be represented as before:

$$\tilde{h}(x,t) = \int_0^t -2k_x(x, t-s) f(e^N s)\, ds \tag{19_N}$$

where

$$k(x,t) = 1/\sqrt{4\pi t}\, e^{-x^2/4t}.$$

By Theorem 4.1 and (19_N),

$$h(x,t) \geqq \tilde{h}(x,t) = \int_0^t -2k_x(x, t-s) f(e^N s)\, ds \tag{32}$$

on π_∞.

From (30) and (32), we have

$$h(x,t) \geqq \hat{h}(x,t) = \frac{c_3 x}{2\sqrt{\pi}} \int_{2K'e^{-N}}^t (t-s)^{-3/2} \exp\left(\frac{-x^2}{4(t-s)}\right) e^{N/2} s^{1/2} \exp\left(c_4 e^{N/2} s^{1/2}\right) ds$$

$$\geqq \hat{h}(1,t) = \frac{c_3}{2\sqrt{\pi}} \exp\left(\frac{N}{2}\right) \int_{(2K'e^{-N})/t}^1 s^{1/2}(1-s)^{-3/2} \cdot \exp\left(c_4 e^{N/2}(ts)^{1/2} - \frac{1}{4(1-s)t}\right) ds$$

for $x \in [0,1]$, where $\hat{h}$ solves $\hat{h}_t = \hat{h}_{xx}$, $\hat{h}(0,t) = c_3 t^{1/2} e^{N/2} \exp\left(c_4 e^{N/2} t^{1/2}\right)$, $\hat{h}(x,0) = 0$ and hence is monotone decreasing in x.

In particular, we may take $t = A e^{-N/3}$ for some positive A. Certainly the requirement that $t > 2K' e^{-N}$ is satisfied for large enough N. Then

$$h(x,t) \geqq \frac{c_3}{2\sqrt{\pi}} e^{N/2} \int_{(2K'e^{-2N/3})/A}^1 s^{1/2}(1-s)^{-3/2} \exp\left[c_4 e^{N/3} A^{1/2} s^{1/2} - \frac{e^{N/3}}{4A(1-s)}\right] ds.$$

The dominant contribution to this integral is from near $s = S$, where the argument of exponential $I = e^{N/3}[c_4 A^{1/2} s^{1/2} - (1/4A(1-s))]$ takes its maximum value. Now I is maximal if

$$\frac{1}{2} c_4 A^{1/2} S^{-1/2} = \frac{1}{4A(1-S)^2},$$

i.e., A and S are related by

$$2c_4 A^{3/2} = \frac{S^{1/2}}{(1-S)^2}$$

or

$$A = (2c_4)^{-2/3} S^{1/3} (1-S)^{-4/3}, \qquad 0 < S < 1. \tag{33}$$

In this case,

$$I = \left[c_4 A^{1/2} S^{1/2} - \frac{1}{4A(1-S)} \right] e^{N/3}$$
$$= 2^{-4/3} c_4^{2/3} S^{-1/3} (1-S)^{-2/3} (3S-1)\, e^{N/3}.$$

Thus, taking A to be given by (33) for some $S \in (\frac{1}{3}, 1)$ we see that the argument of the integral near $s = S$ is itself exponentially large, $O(e^{N/3})$, and $h(x, t)$ must exceed 1 for $t \geqq A\, e^{-N/3}$.

In particular, for N sufficiently large, $\hat{h}(x, t) \geqq 1$ and $h(x, t) = h_N(x, t) \geqq 1$ for $t \geqq \tau_N$ with $\tau_N = A\, e^{-N/3}$. Hence,

$$h_N(x, t_N) \geqq 1$$

for all $x \in \Omega$, where $t_N \leqq \tau_N = A\, e^{-N/3}$.

But $\tau_N = A\, e^{-N/3}$ implies $\sum_N \tau_N < \infty$, which in turn implies $T < \infty$. This contradicts our original assumption and completes the proof of Theorem 2.1.

REFERENCES

[1] J. BEBERNES, A. BRESSAN, AND A. LACEY, *Total blowup versus single point blowup*, J. Differential Equations, 73 (1988), pp. 30–44.

[2] J. BEBERNES AND W. FULKS, *The small heat-loss problem*, J. Differential Equations, 57 (1985), pp. 324–332.

[3] J. BEBERNES AND D. KASSOY, *A mathematical analysis for blow-up for thermal reactions*, SIAM J. Appl. Math., 40 (1981), pp. 476–484.

[4] J. G. BURNELL, A. A. LACEY, AND G. C. WAKE, *Steady states of reaction-diffusion equations. Part* I: *Questions of existence and continuity of solution branches*, J. Austral. Math. Soc. Ser. B, 24 (1983), pp. 374–391.

[5] A. FRIEDMAN AND B. MCLEOD, *Blowup of positive solutions of semilinear heat equations*, Indiana Univ. Math. J., 34 (1985), pp. 425–447.

[6] A. FRIEDMAN AND Y. GIGA, *A single point blowup for solutions of semilinear parabolic systems*, J. Fac. Sci. Univ. Tokyo, Sect. 1A Math., 34 (1987), pp. 65–79.

[7] S. L. HOLLIS, R. H. MARTIN, AND M. PIERRE, *Global existence and boundedness in reaction-diffusion systems*, SIAM J. Math. Anal., 18 (1987), pp. 744–761.

[8] A. A. LACEY, *Mathematical analysis of thermal runaway for spatially inhomogeneous reactions*, SIAM J. Appl. Math., 43 (1983), pp. 1350–1366.

[9] A. LACEY AND D. TZANETIS, *Global existence and convergence to a singular steady state for a semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 105 (1987), pp. 289–305.

[10] C. V. PAO, *Asymptotic behavior and nonexistence of global solutions for a class on nonlinear boundary value problems of parabolic type*, J. Math. Anal. Appl., 65 (1978), pp. 616–637.

[11] ———, *On nonlinear reaction-diffusion systems*, J. Math. Anal. Appl., 87 (1982), pp. 165–198.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1426-1452, November 1990
© 1990 Society for Industrial and Applied Mathematics
005

A DIRICHLET PROBLEM EXHIBITING GLOBAL BIFURCATION WITH SYMMETRY BREAKING*

CHRISTOPH POSPIECH†

Abstract. A theorem is applied to a semilinear boundary value problem in order to establish the existence of both a branch of radially symmetric solutions and branches with less symmetry bifurcating from this set. The latter branches are proved to satisfy a Rabinowitz-type alternative.

Key words. global bifurcation, symmetry breaking, nonlinear functional analysis, index theory, fixed point theorems, partial differential equations

AMS(MOS) subject classifications. 35B32, 58G10

Introduction. Bifurcation problems frequently arise in all areas of science and engineering. One example is elasticity theory and structural mechanics, where there are usually symmetries inherent to the problems. Here it is easier to look for the symmetric solutions first, while the solutions having less symmetry may be found when they bifurcate from the previous ones. In this paper we restrict our attention to a partial differential equation (PDE) serving as a model problem for the more complicated equations in structural mechanics.

The main theorem presented in § 1 of this article is a generalisation of a result of Rabinowitz [10] adapted to the needs of the following sections. In particular we had to pose fairly mild assumptions on the primary branch calling for a revised version of the proof of Rabinowitz, which is therefore included. We also draw the connection to local bifurcation theory by formulating a corollary, which generalises in a straightforward way the well-known local bifurcation theorem of Crandall and Rabinowitz [5, Thm. 1, p. 323] and the local bifurcation results of Cicogna [3, Thm. 2, p. 790] and Vanderbauwhede, [16, p. 205]. Thus it is not surprising that we get bifurcation. The main point is that *any local bifurcation obtained by application of any of the above results already is global in the sense that it satisfies a Rabinowitz-type alternative. No extra assumption is needed except that the nonlinear map F satisfies the compactness assumption below.*

In § 2 the general bifurcation theorem is applied to establish the existence of a branch of radially symmetric solutions for a semilinear PDE with Dirichlet boundary conditions. We consider only nonlinearities with linear growth at infinity having a single zero. Here the branch of radially symmetric solutions can be proven to follow a slalom course around some flags set up by the radially symmetric solutions of the corresponding Neumann problem. Along the branch the type of solution changes from positive (or negative) solutions to Mexican hat shaped (one nodal line), and then the solutions successively take up nodal lines without reaching a bound for their number. Previously, existence was known only for that part of the branch containing positive (or negative) solutions [4], [8], [13]. Our method, however, has the disadvantage of yielding a branch of solutions, which is only known to be a closed connected set, not necessarily a smooth curve.

* Received by the editors December 27, 1988; accepted for publication (in revised form) November 20, 1989. This article was part of the author's doctoral dissertation at Ruprecht-Karls-Universität, Heidelberg, 1988.

† IBM Scientific Center, Tiergartenstrasse 15, D-6900, Heidelberg, Federal Republic of Germany.

In the final section this closed connected set, being the output of the abstract bifurcation theorem, is fed into the theorem again to prove bifurcation of less symmetric solutions. The existence of these bifurcations depends on a condition that is independent of the nonlinearity and can be calculated explicitly. (Some tables are appended at the end of this paper.) The bifurcating branches satisfy a Rabinowitz-type alternative.

Smoller and Wasserman [14] have already found a bifurcation from that part of the branch containing positive solutions. This bifurcation was later proven to be global by Cerami [1]. By table-lookup we also can assure bifurcation from that part of the branch containing positive or Mexican hat shaped solutions. But since the symmetries along the bifurcating branches differ in both cases and since the solutions found by Smoller are the only ones bifurcating at this point, we must have found a bifurcation different from Smoller's. Smoller also proved that his bifurcation point is the only one for positive solutions. Thus ours cannot be a positive solution, rather it must be Mexican hat shaped. Further table-lookup reveals many additional bifurcations from solutions with several nodal lines.

We also give a closer look at Smoller's bifurcation. Both Smoller and Cerami used a transversality condition to establish their results. We will prove that the transversality condition is automatically satisfied and can be dropped.

1. An abstract bifurcation theorem.

1.1. Setup and results.

General assumptions. In this paper we will use the following notation:

(1) X, Y denote Banach spaces. G denotes some group operating on both X and Y by bounded linear operators, i.e., there are group homomorphisms $G \to B(X, X)$ and $G \to B(Y, Y)$. For simplicity we will write gx instead of $[\varphi(g)](x)$, where φ is one of the group homomorphisms above.

(2) There is some nonlinear mapping $F: \mathbb{R} \times X \to Y$ of the following form:

$$F(\lambda, u) = Bu + \Phi(\lambda, u),$$

where $B \in B(X, Y)$ is a bounded linear operator with inverse $B^{-1} \in B(Y, X)$, and $\Phi \in C^0(\mathbb{R} \times X, Y)$ is continuous and compact. If there is a compact equivariant embedding $\iota: X \to Y$, we can replace Φ by $\tilde{\Phi} = \Phi \circ (\mathrm{Id}, \iota)$, where $\Phi \in C^0(\mathbb{R} \times Y, Y)$ no longer needs to be compact.

(3) Both B and Φ are equivariant (i.e., $B(gu) = gBu$ and $\Phi(\lambda, gu) = g\Phi(\lambda, u)$ for all $g \in G$, $\lambda \in \mathbb{R}$, $u \in X$).

Definitions. The term *symmetry* is used as a synonym for the isotropy subgroup

$$G_u := \{g \in G | gu = u\}$$

of some element $u \in X$ (respectively, $u \in Y$). Given some subgroup $H \subseteq G$, the *space of minimum symmetry* H is given by the closed subspace

$$X^H = \{u \in X | H \subseteq G_u\} = \{u \in X | hu = u \text{ for all } h \in H\},$$

which is sometimes also called *isotropy subspace* of H. Because of the equivariance of F we have $F(\mathbb{R} \times X^H) \subseteq Y^H$. Thus we can define the restriction

$$F^H := F|_{\mathbb{R} \times X^H}: \mathbb{R} \times X^H \to Y^H.$$

A zero (λ, u) of F^H is called

- *H-regular*, if the (Fréchet-) derivative DF^H is onto in a neighbourhood of (λ, u) and depends continuously on (λ, u).[1]

[1] This is just enough to apply the implicit function theorem.

• *Interior point* of some subset M of zeros of F^H, if, in a neighbourhood of (λ, u) in X^H, all zeros of F^H are contained in M.

• *Special boundary point* of some subset M of zeros of F^H, if, in a neighbourhood of (λ, u) in X^H, the zeros of F^H are given by some curve $t \mapsto (\lambda + \alpha t, u(t))$, $\alpha \in \{\pm 1\}$ satisfying

$$u(0) = u,$$

$$(\lambda + \alpha t, u(t)) \in M \quad \text{if } t \leqq 0,$$

$$(\lambda + \alpha t, u(t)) \notin M \quad \text{if } t > 0.$$

Definition (of the *H*-index). For every isolated zero u of $F^H(\lambda, \cdot)$ and every isolating neighbourhood $U(u) \subseteq X^H$ we define the *H-index* of (λ, u) to be the following:

$$i_H(\lambda, u) := \deg((B^{-1} \circ F)^H(\lambda, \cdot)|_{U(u)}, 0).$$

The expression on the right-hand side denotes the Leray-Schauder degree (with respect to zero) of $(B^{-1} \circ F)^H(\lambda, \cdot) = \mathrm{Id}_{X^H} + (B^{-1} \circ \Phi)^H(\lambda, \cdot)$ restricted to the neighbourhood $U(u)$. This does not depend on the choice of $U(u)$ (Eisenack and Fenske [6, p. 119]).

The main theorem presented now is a generalisation of a result of Rabinowitz [10] to the situation where some symmetry group is present and bifurcation from a closed bounded set of solutions is considered. We care for the symmetry by picking a subgroup and restricting the problem to the invariant subspace corresponding to that subgroup. The general form of primary branch is needed because in § 3 we apply this theorem to prove bifurcation from a closed connected set of radially symmetric solutions, the existence of which was established in § 2 by the same theorem. Thus this branch of radially symmetric solutions may be almost arbitrarily wild. In this respect the theorem also generalises results of Schaaf [11, Thm. 2.3, p. 13] and Cicogna [2]. See Fig. 1 for a general bifurcation diagram.

THEOREM 1.1. (a) *Suppose we have the following*:

(1) $\mathcal{S}_1$ *denotes a closed bounded set of zeros* (λ, u) *of* F^H, *where* λ *is taken from some interval* $[a, b] \subset \mathbb{R}$.

(2) *All zeros lying above the interval boundaries*[2] *are special boundary points of* $\mathcal{S}_1$.

(3) *Finally, the H-indices taken over zeros lying above the left and right interval boundary, respectively, add up to different sums*:

$$(1) \qquad \sum_{(a,u) \in \mathcal{S}_1} i_H(a, u) \neq \sum_{(b,u) \in \mathcal{S}_1} i_H(b, u).$$

Then, above the interval interior, $\mathcal{S}_1$ *intersects some continuum* $\mathcal{C}$ *of zeros of* F^H. *The bifurcating branch* $\mathcal{C}$ *either is unbounded (as a subset of* $\mathbb{R} \times X$*) or constitutes a continuation of* $\mathcal{S}_1$ *across the above-mentioned special boundary point.*[3]

(b) *Suppose further that* $\mathcal{S}_1$ *is part of a closed set* $\mathcal{S}_0$ *of zeros of* F^H, *which contains every special boundary point* (a, u), $(b, u) \in \mathcal{S}_1$ *of* $\mathcal{S}_1$ *as interior point. Moreover, these special boundary points are the only elements of the intersection of* $\mathcal{S}_1$ *and the closure of* $\mathcal{S}_0 \setminus \mathcal{S}_1$. *Then, by passing to a subset, the continuum* $\mathcal{C}$ *satisfies the following*:

- $\mathcal{C}$ *and* $\mathcal{S}_0$ *share no interior point (of* $\mathcal{S}_0$*),*
- $\mathcal{C}$ *intersects* $\mathcal{S}_1$,
- $\mathcal{C}$ *is unbounded (as a subset of* $\mathbb{R} \times X$*) or intersects* $\mathcal{S}_0 \setminus \mathcal{S}_1$.

[2] That is, those zeros (λ, u), where $\lambda = a$ or $\lambda = b$.

[3] That is, in a small neighbourhood of the special boundary point, every zero of F^H either is in $\mathcal{C}$ or in $\mathcal{S}_1$.

In the above theorem the primary branch $\mathcal{S}_1$ may be almost arbitrarily wild. If, however, this branch is a curve (at least locally), the other assumptions can be substantially simplified. This is formulated as a corollary, which generalises in a straightforward way the well-known local bifurcation theorem of Crandall and Rabinowitz [5, Thm. 1, p. 323] and the local bifurcation results of Cicogna [3, Thm. 2, p. 790] and Vanderbauwhede [16, p. 205]. Thus it is not surprising that we get bifurcation. The main point is that *any local bifurcation obtained by application of any of the above results already is global in the sense that it satisfies a Rabinowitz-type alternative. No extra assumption is needed except that the nonlinear map F satisfies the general assumptions above.* In particular the Leray-Schauder degree arguments being adopted here rely on information about the algebraic multiplicities of eigenvalues, while in the following we only make assumptions on the kernel, i.e., on the geometric multiplicity of the eigenvalue 0. The main part of the proof will be Lemma 1.7, which bridges this gap—without assuming (semi-)simplicity for the eigenvalue 0.

COROLLARY 1.2. *Let $\mathcal{S}_0$ be a closed set of zeros of F^H and assume that sufficiently close to some point $(\lambda_0, u_0) \in \mathcal{S}_0$ the map F^H is C^2 and the set $\mathcal{S}_0$ can be parametrized by a C^1 curve $t \mapsto (\lambda(t), u(t))$ with $(\lambda(0), u(0)) = (\lambda_0, u_0)$ and with tangent vector $\bar{v} := (\lambda'(0), u'(0))$. Assume further that*

(1) *$\mathbb{R} \cdot \bar{v}$ has a complement W^H in the kernel of $DF^H(\lambda_0, u_0)$,*

$$\ker DF^H(\lambda_0, u_0) = \mathbb{R} \cdot \bar{v} \oplus W^H,$$

and W^H has odd dimension.

(2) *(Transversality.) For all $w \in W^H \setminus \{0\}$ we have*

$$D^2F^H(\lambda_0, u_0)(\bar{v}, w) \notin \operatorname{im} DF^H(\lambda_0, u_0).$$

Then the following statements hold:

(1) *In a sufficiently small neighbourhood of (λ_0, u_0) all zeros of F^H except (λ_0, u_0) are H-regular and thus are interior points of $\mathcal{S}_0$.*

(2) *There is a continuum $\mathcal{C}$ of zeros of F^H such that*

- *$\mathcal{C}$ and $\mathcal{S}_0$ share no interior point of $\mathcal{S}_0$,*
- *$\mathcal{C}$ contains (λ_0, u_0),*
- *$\mathcal{C}$ is unbounded or intersects with $\mathcal{S}_0$ in a zero different from (λ_0, u_0).*

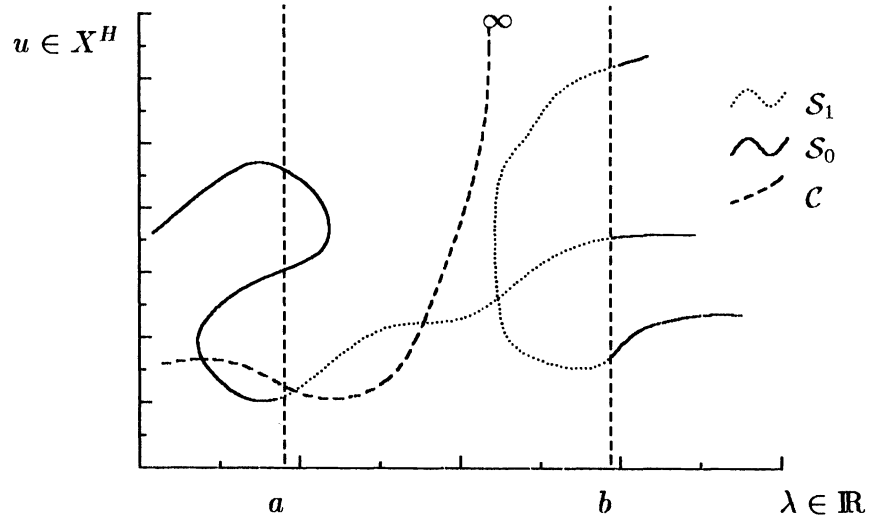


FIG. 1. *A general bifurcation diagram.*

Similar to the paper of Crandall and Rabinowitz [5] we can reduce this general result to the case of the curve being parametrized over λ by choosing an appropriate transformation. This time, however, this transformation should be defined globally and should take care of the symmetry. To achieve this, let $\phi \in (\mathbb{R} \times X)^*$ be a functional on $\mathbb{R} \times X$ with $\langle \phi, \bar{v} \rangle = 1$ and $\langle \phi, W^H \rangle = 0$. We now introduce the spaces $\hat{X} := \mathbb{R} \times X^H$ and $\hat{Y} := \mathbb{R} \times Y^H$, and a new curve parameter $\rho := \rho(t) := \langle \phi, (\lambda(t), u(t)) \rangle$. We then define

$$\hat{F} : \mathbb{R} \times \hat{X} \mapsto \hat{Y}$$

by

$$\begin{aligned} \hat{F}(\rho, \lambda, u) &:= (\rho - \langle \phi, (\lambda, u) \rangle, F(\lambda, u)) \\ &= (\lambda, Bu) + (\rho - \lambda - \langle \phi, (\lambda, u) \rangle, \Phi(\lambda, u)). \end{aligned}$$

This map satisfies the general assumptions above, provided G operates trivially on $\hat{X}$ and $\hat{Y}$, respectively. $\hat{F}$ also vanishes by construction along the curve $\rho \mapsto (\lambda(\rho), u(\rho))$. So the corollary above can be obtained by applying the following to $\hat{F}(\rho, \hat{u}) = \hat{F}(\rho, \lambda, u)$.

COROLLARY 1.3. *Let $\mathcal{S}_0$ be a closed set of zeros of F^H and assume that sufficiently close to some point $(\lambda_0, u_0) \in \mathcal{S}_0$ the map F^H is C^2 and the set $\mathcal{S}_0$ can be parametrized by a C^1 curve $\lambda \mapsto (\lambda, u(\lambda))$. Assume further that $L^H(\lambda)$, defined by*

$$L^H(\lambda) := D_u F^H(\lambda, u(\lambda)) \in B(X^H, Y^H),$$

satisfies the following:

(1) *The kernel $W^H := \ker L^H(\lambda_0)$ of $L^H(\lambda_0)$ has odd dimension.*

(2) (*Transversality.*) *For all $0 \neq w \in W^H$ we have*

$$D_\lambda L^H(\lambda_0) w \notin \operatorname{im} L^H(\lambda_0).$$

Then the statements of the previous corollary hold.

1.2. Proofs of the abstract results. For the rest of the section $\mathcal{S}_1$ is assumed to satisfy (a)(1) and (a)(2) of Theorem 1.1.

DEFINITION. For every special boundary point (λ, u) of $\mathcal{S}_1$ lying above the boundaries of the interval we can find an open neighbourhood $U(\lambda, u) \subseteq \mathbb{R} \times X^H$, such that the zeros of F^H in $\overline{U(\lambda, u)}$ form a continuous curve $t \mapsto (\lambda + \alpha t, u(t))$, where $\alpha \in \{\pm 1\}$.

We keep these neighbourhoods for the rest of this section and denote their union by

$$U := \bigcup_{\substack{\lambda \in \{a,b\} \\ (\lambda,u) \in \mathcal{S}_1}} U(\lambda, u).$$

$\tilde{U}$ is the subset of U containing the elements with $\lambda \in]a, b[$.

LEMMA 1.4. *Suppose there is a neighbourhood $V \subseteq \mathbb{R} \times X^H$ of $\mathcal{S}_1 \setminus U$, such that every zero of F^H lying on the boundary ∂V of V is already contained in $\tilde{U}$. Then we have*

$$\sum_{(a,u) \in \mathcal{S}_1} i_H(a, u) = \sum_{(b,u) \in \mathcal{S}_1} i_H(b, u).$$

Proof. Choose $\lambda_1, \lambda_2 \in \mathbb{R}$ as suggested in Fig. 2, such that $\bar{V} \subseteq]\lambda_1, \lambda_2[\times X^H$. By assumption there are neither zeros (λ, u) of F^H on the boundary ∂V of V for

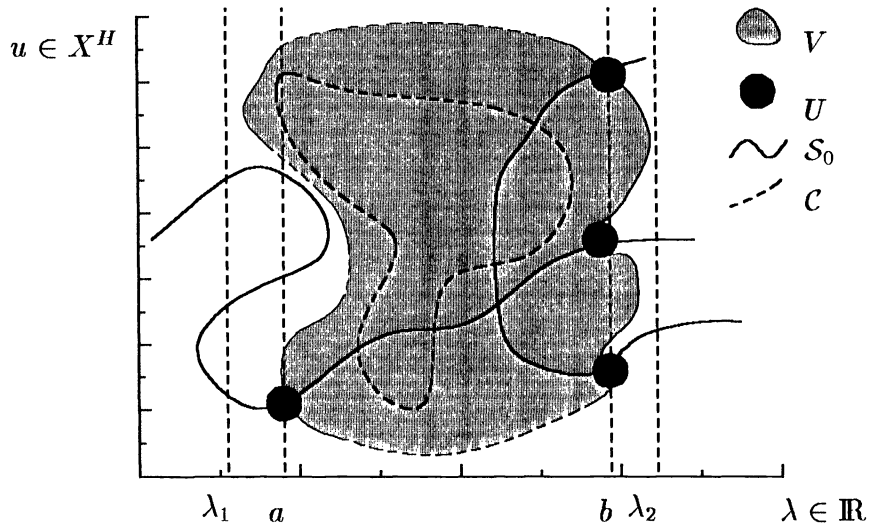


FIG. 2. *The open sets V and U.*

$\lambda \in [\lambda_1, a] \cup [b, \lambda_2]$ nor on the boundary $\partial(U \cup V) = (\partial U \setminus V) \cup (\partial V \setminus U)$ of $U \cup V$ for $\lambda \in [a, b]$. So the assertion follows by homotopy invariance of the Leray-Schauder degree of $(B^{-1} \circ F)^H(\lambda, \cdot)$. □

The following lemma of Whyburn [18, Cor. 9.3] is quoted without proof.

LEMMA 1.5. *If A and B are disjoint closed subsets of a compact metric space such that no component of K intersects both A and B, there exists a separation $K = K_A \cup K_B$, where K_A and K_B are disjoint compact sets containing A and B, respectively.*

COROLLARY 1.6. *Let X be a compact metric space, N be a closed, connected subset, and $V \subseteq X$ be a subset satisfying $N \cap V \neq \emptyset \neq N \setminus V$. Then the closure of every component of $N \cap V$ meets the boundary ∂V of V. Similarly, the closure of every component of $N \setminus V$ meets the boundary ∂V of V.*

A similar corollary can be found in Whyburn [18, Lemma 10.1], so we again skip the proof.

Proof of Theorem 1.1. Since we will apply the Whyburn alternative (Lemma 1.5), we need a compact metric space with which to work. We therefore construct the Alexandroff compactification $\mathcal{S} = (F^H)^{-1}(0) \cup \{\infty\}$ of the zero set of F^H. Because of Φ^H being compact, this zero set is locally compact. Thus the compactification exists, is second countable, and therefore metrizable.

(a) Let $A := \mathcal{S}_1 \setminus U$ and B consist of ∞ and the special boundary points of $\mathcal{S}_1$ that lie above the interval boundaries. We will prove that some component of $K := \mathcal{S} \setminus \tilde{U}$ meets both A and B.

Suppose this is not so. Since both A and B are closed in $\mathcal{S}$ and thus in K, the Whyburn alternative asserts the existence of a separation $K = K_A \cup K_B$, where $K_A \supset A$ and $K_B \supset B \ni \infty$ are disjoint compact subsets of K. Take

$$\delta := \tfrac{1}{2} \operatorname{dist}(K_B \setminus \{\infty\}, K_A) > 0,$$

where "dist" refers to the norm in $\mathbb{R} \times X^H$.[4] Then $V := U_\delta(K_A)$ meets all requirements

[4] If $K_B \setminus \{\infty\} = \emptyset$, take $\delta = 1$.

of Lemma 1.4 and a contradiction to assumption (a)(3) on the nonequality of index sums follows immediately.

(b) Let $\mathscr{C}_2$ be the component of $K = \mathscr{S} \setminus \tilde{U}$ just obtained. Applying Corollary 1.6 to the metric space K, the connected set $\mathscr{C}_2$, and $A \subseteq K$, we find that the closure of every component of $\mathscr{C}_2 \setminus A$ meets the boundary ∂A of A (closure and boundary are taken with respect to K). The boundary ∂A does not contain any interior point of $\mathscr{S}_0$. Otherwise, if all zeros of F^H in some neighbourhood $\mathscr{U}(x)$, $x \in \partial A$ were contained in $\mathscr{S}_0$, we would have $\mathscr{U}(x) \cap \overline{\mathscr{S}_0 \setminus \mathscr{S}_1} = \varnothing$ after passing to a subset. Thus $\mathscr{U}(x) \cap K \subseteq A$, which contradicts $x \in \partial A$.

Now let $\mathscr{C}_1$ be the closure of one of those components of $\mathscr{C}_2 \setminus A$ that meet B. We repeat the above argument with $\mathscr{C}_2$ being replaced by $\mathscr{C}_1$ and A being replaced by $\tilde{A} := (\mathscr{S}_0 \setminus \mathscr{S}_1) \cup B$. Again the closure (with respect to K) of every component of $\mathscr{C}_1 \setminus \tilde{A}$ meets the boundary $\partial \tilde{A}$ of $\tilde{A}$, which does not contain any interior point of $\mathscr{S}_0$. Otherwise, if all zeros of F^H in some neighbourhood $\mathscr{U}(x)$, $x \in \partial \tilde{A}$ were contained in $\mathscr{S}_0$, we would have $\mathscr{U}(x) \cap A = \varnothing$ after passing to a subset. Thus $\mathscr{U}(x) \cap K \subseteq \tilde{A}$, which contradicts $x \in \partial \tilde{A}$.

To get the bifurcating branch $\mathscr{C}$, we select a component of $\mathscr{C}_1 \setminus \tilde{A}$ intersecting A and take its closure with respect to $\mathbb{R} \times X^H$. By construction $\mathscr{C}$ is either unbounded or closed in K. In the latter case $\mathscr{C}$ contains a point in $\mathscr{S}_0 \setminus \mathscr{S}_1$, because all zeros in $B \setminus \{\infty\}$ are interior points of $\mathscr{S}_0$. □

The following lemma is the key ingredient of the proof of Corollary 1.3. It bridges the gap between the assumptions to be made for local and global bifurcation theory.

LEMMA 1.7. *Given a C^1-function $\lambda \mapsto L(\lambda) \in B(X, X)$ with $L(\lambda) - \mathrm{Id}$ compact for every $\lambda \in \mathbb{R}$ and* $\ker L(\lambda_0) \neq \{0\}$, *let $\Gamma_1 \subseteq \mathbb{C}$ be a positively oriented Jordan curve that separates the eigenvalue 0 from the rest of the spectrum of $L(\lambda_0)$. Define*

$$\det_{\Gamma_1}(L(\lambda)) \quad \text{for small } |\lambda - \lambda_0|$$

to be the product of the eigenvalues (counted with algebraic multiplicity) that are enclosed by Γ_1.

Then $\lambda \mapsto \det_{\Gamma_1}(L(\lambda))$ is a real C^1-function (for $|\lambda - \lambda_0|$ small) satisfying

$$\det_{\Gamma_1}(L(\lambda)) = c \cdot (\lambda - \lambda_0)^m + o((\lambda - \lambda_0)^m),$$

where $m = \dim \ker L(\lambda_0)$ and $c \neq 0$ if and only if

$$L'(\lambda_0) w \notin \operatorname{im} L(\lambda_0) \quad \text{for all } 0 \neq w \in \ker L(\lambda_0).$$

Proof. We define the projection $P_1(\lambda)$ by

$$P_1(\lambda) := -\frac{1}{2\pi i} \int_{\Gamma_1} (L(\lambda) - \zeta)^{-1} \, d\zeta \in B(X, X).$$

The restriction of $L(\lambda)$ to the finite-dimensional space $\operatorname{im} P_1(\lambda)$ has a spectrum consisting only of those eigenvalues of $L(\lambda)$ that are enclosed by Γ_1 (Kato [7, Thm. 6.17, p. 178]). Let $e_1, \cdots, e_n$ be a basis of $\operatorname{im} P_1(\lambda_0)$, such that $e_1, \cdots, e_m$ span the kernel of $L(\lambda_0)$ and their exterior product satisfies $e_1 \wedge \cdots \wedge e_n = 1$. Now we define $l_j(\lambda) := T^{-1}(\lambda) L(\lambda) T(\lambda) e_j$, where

$$T(\lambda) := P_1(\lambda) P_1(\lambda_0) + (1 - P_1(\lambda))(1 - P_1(\lambda_0)) \in B(X, X).$$

Because of $T(\lambda_0) = \mathrm{Id}_X$ the map $T(\lambda)$ is an isomorphism on X for small $|\lambda - \lambda_0|$ that

takes Bild $P_1(\lambda_0)$ onto Bild $P_1(\lambda)$. Thus we get

$$\begin{aligned}
\det_{\Gamma_1}(L(\lambda)) &= \det\left(L(\lambda)|_{\operatorname{im} P_1(\lambda)}\right) \\
&= \bigwedge_{j=1}^{n} l_j(\lambda) \\
&= \bigwedge_{j=1}^{n} \left(l_j(\lambda_0) + l_j'(\lambda_0)(\lambda-\lambda_0) + o(\lambda-\lambda_0)\right) \\
&= \left(\bigwedge_{j=1}^{m} l_j'(\lambda_0) \wedge \bigwedge_{j=m+1}^{n} l_j(\lambda_0)\right)(\lambda-\lambda_0)^m + o((\lambda-\lambda_0)^m) \\
&= \underbrace{\left(\bigwedge_{j=1}^{m} L'(\lambda_0)e_j \wedge \bigwedge_{j=m+1}^{n} L(\lambda_0)e_j\right)}_{=:\,c}(\lambda-\lambda_0)^m + o((\lambda-\lambda_0)^m).
\end{aligned}$$

The last equation follows from

$$l_j'(\lambda_0) = L'(\lambda_0)e_j + L(\lambda_0)T'(\lambda_0)e_j,$$

since the exterior product vanishes, provided its factors are linearly dependent. For the same reason we have $c \neq 0$, if and only if $L'(\lambda_0)w \notin \operatorname{im} L(\lambda_0)$ for all nontrivial $w \in \ker L(\lambda_0)$. □

Proof of Corollary 1.3. We construct a set $\mathcal{S}_1 \subseteq \mathcal{S}_0$ and check that it meets the requirements of Theorem 1.1. Without loss of generality we can assume that $X = Y$ and $B = \mathrm{Id}_X$, because replacing F by $B^{-1}F$ neither affects the result nor the assumptions of Corollary 1.3.

Remember that in some neighbourhood $[a, b] \times W \subseteq \mathbb{R} \times X^H$ of (λ_0, u_0) the set $\mathcal{S}_0$ can be parametrized by a C^1-curve $\varphi(\lambda) = (\lambda, u(\lambda))$. We choose $\mathcal{S}_1 := \varphi([a, b])$ to be the image of the compact interval $[a, b]$. If $[a, b]$ is chosen small enough, every point of this image is H-regular except for (λ_0, u_0), since Lemma 1.7 implies that

$$\det_{\Gamma_1}(L^H(\lambda)) = c \cdot (\lambda-\lambda_0)^m + o((\lambda-\lambda_0)^m) \neq 0 \quad \text{for } \lambda \neq \lambda_0, |\lambda-\lambda_0| \text{ small.}$$

It mainly remains to compare the H-indices at the special boundary points $\varphi(a)$ and $\varphi(b)$ of $\mathcal{S}_1$. For $\lambda_0 \neq \lambda \in [a, b]$ the H-index of $(\lambda, u(\lambda))$ is given by

$$i_H(\lambda, u(\lambda)) = \operatorname{sign} L^H(\lambda) = (-1)^\beta, \tag{2}$$

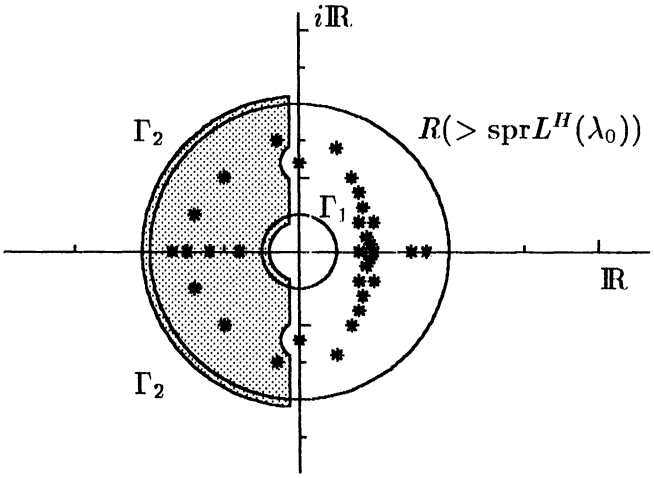


FIG. 3. *The * indicate the spectrum of* $L^H(\lambda_0)$.

where β is the number of real negative eigenvalues of $L^H(\lambda)$ counted with their algebraic multiplicity (Eisenack and Fenske [6, Thm. 5.4.4, No. 12, p. 120 and definition A.1.14, p. 221]). Since $L^H(\lambda)$ is a map on real Banach spaces, eigenvalues with nonzero imaginary part always exist in conjugacy pairs (with equal algebraic multiplicity). So these eigenvalues do not contribute to the degree and we can write

$$\begin{aligned} i_H(\lambda, u) &= (-1)^{\dim \operatorname{im} P_2(\lambda)} \cdot \operatorname{sign} \det_{\Gamma_1}(L^H(\lambda)) \\ &= (-1)^{\dim \operatorname{im} P_2(\lambda)} \cdot \operatorname{sign}(c) \cdot \operatorname{sign}(\lambda - \lambda_0)^m, \end{aligned} \tag{3}$$

where the projection $P_2(\lambda)$ is defined by

$$P_2(\lambda) = -\frac{1}{2\pi i} \int_{\Gamma_2} (L(\lambda) - \zeta)^{-1} \, d\zeta$$

and the positively oriented Jordan curve $\Gamma_2 \in \mathbb{C}$ is chosen according to Fig. 3. Since the spaces im $P_2(\lambda)$ and im $P_2(\lambda_0)$ are isomorphic for every $\lambda \in [a, b]$ (Kato [7, Lemma 4.10, p. 34 and footnote on p. 33]), the first factor on the right-hand side of equation (3) is independent of $\lambda \in [a, b]$. So, up to some constant ± 1, the H-index of $(\lambda, u(\lambda))$ is just sign $(\lambda - \lambda_0)^m$ and thus changes sign at λ_0, because $m = \dim \operatorname{Kern} L^H(\lambda_0)$ is odd by assumption. □

2. A radial slalom course.

2.1. Preparation of the course. We now apply Theorem 1.1 to the elliptic PDE

$$\Delta u + f(u) = 0 \quad \text{in } \mathbb{B}_R(0), \tag{4}$$

where $\mathbb{B}_R(0) \subset \mathbb{R}^n$ denotes a ball of dimension $n \geqq 2$ with radius R and center 0. The problem has an inherent $O(n)$-symmetry, since the domain $\mathbb{B}_R(0)$ is invariant under rotations and reflections. The investigation of radially symmetric solutions reduces the PDE to an ordinary differential equation (ODE) boundary value problem. For this reduced problem we consider both Dirichlet and Neumann boundary conditions, since the solutions of the Neumann problem turn out to be the key in understanding the Dirichlet problem.

Artificially introducing some initial value we arrive at the following initial value problem:

$$\begin{aligned} &u'' + \frac{n-1}{r} u' + f(u) = 0 \quad \text{in } \mathbb{R}^+ \\ &u(0) = p. \end{aligned} \tag{5}$$

General assumptions. f is assumed to be *admissible*, which means the following:
(1) f has precisely one zero, which is regular.
(2) f is continuously differentiable and satisfies

$$\lim_{u \to \infty} f'(u) = \lim_{u \to -\infty} f'(u) =: a > 0. \tag{6}$$

Applying l'Hôpital's rule we derive

$$\lim_{u \to \pm\infty} \frac{f(u)}{u} = a. \tag{7}$$

Notation. Since f grows linearly at infinity we would like to include the solution $u(r, p)$ of (5) for $p = \pm\infty$. We therefore define an angular variable $\vartheta = \arctan p$ and

consider the solution $w(r, \vartheta) := \cos \vartheta \cdot u(r, \tan \vartheta)$ of the initial value problem

$$w'' + \frac{n-1}{r} w' + \cos \vartheta \cdot f\left(\frac{w}{\cos \vartheta}\right) = 0 \quad \text{in } \mathbb{R}^+, \tag{8}$$
$$w(0) = \sin \vartheta.$$

By ϑ_0 we denote the solution of $f(\tan \vartheta_0) = 0$ satisfying $|\vartheta_0| < \pi/2$.

DEFINITION. Given some real function J, the positive zeros of which are discrete, the kth zero of J will be denoted by $[J]_k \in \mathbb{R}$.

THEOREM 2.1 (Neumann problem). *Apart from the trivial solution*

$$\{(r, \vartheta_0 + m\pi) | r \geqq 0, m \in \mathbb{Z}\}$$

the zeros of $w_r(r, \vartheta)$ can be parametrized by continuous functions

$$R_k(\vartheta) := [w_r(\cdot, \vartheta)]_k.$$

All of these functions have period π and are therefore bounded.

THEOREM 2.2 (Dirichlet problem). *The zeros of $w(r, \vartheta)$ are periodic in ϑ with period π. Depending on the sign of the zero ϑ_0 of $f \circ \tan$ one of the following holds:*

$\vartheta_0 > 0$: *$w(r, \vartheta)$ has a component $\mathscr{C}_0$ of zeros satisfying the following.*

- *$\mathscr{C}_0$ meets $(0, \pi)$ and is unbounded in r.*
- *The component $\mathscr{C}_0$ is a subset of $\mathbb{R} \times]\vartheta_0, \vartheta_0 + \pi[$.*
- *Being unbounded in r the component $\mathscr{C}_0$ must intersect the graphs of the functions R_k describing the Neumann solutions. The crossing points satisfy*

$$\operatorname{sign}(\cos \vartheta) = (-1)^{k+1}.$$

Thus $\mathscr{C}_0$ traces a slalom course around the "flags" $(R_k(\pi/2), \pi/2)$.

- *$\mathscr{C}_0$ intersects $\vartheta = \pi/2$ in $r = [J_\nu]_k/\sqrt{a}$, where J_ν denotes the Bessel function of order $\nu := (n-2)/2$.*

$\vartheta_0 < 0$: *Almost as in the previous case, but now the slalom course starts at $(0, 0)$ and the crossing condition has changed to*

$$\operatorname{sign}(\cos \vartheta) = (-1)^k.$$

Thus the flags are passed by in the opposite way.

$\vartheta_0 = 0$: *The solution set consists of infinitely many branches bifurcating from the trivial solution*

$$\{(r, m\pi) | r \geqq 0, m \in \mathbb{Z}\}$$

at $\hat{r}_k = [J_\nu]_k/\sqrt{f'(0)}$, where $\nu := (n-2)/2$. These branches can be parametrized by continuous functions

$$r_k(\vartheta) := [w(\cdot, \vartheta)]_k$$

satisfying $r_k(\pi/2) = [J_\nu]_k/\sqrt{a}$. All of these functions have period π and are therefore bounded.

Remark. Although the assumption on the admissibility of f is used only in the proof of Lemma 2.3 below, the above statements are false without it.

$w(r, \vartheta)$ and the zeros of w and w_r are shown in Fig. 4 and Fig. 5.

2.2. Zigzagging around some technical details.

Remark. If f is continuous then the function $g(q, w) = qf(w/q)$ can be continuously extended for $q = 0$, if and only if (7) holds. In this case the extension has the form

$$g(q, u) = \begin{cases} qf(u/q) & \text{if } q \neq 0, \\ au & \text{if } q = 0. \end{cases}$$

g defined this way is even locally Lipschitz because of condition (6). So by [13, p. 157] we have local existence and uniqueness of solutions for the initial value problem (8). The solution is bounded and exists for r arbitrarily large [9, p. 29].

LEMMA 2.3. *Let g be admissible and let its zero be denoted by c. Moreover, let $w(r, p)$ be the solution of the initial value problem*

$$w'' + \frac{n-1}{r} w' + g(w) = 0 \quad \text{in } \mathbb{R}^+ \tag{9}$$

$$w(0) = p.$$

Then $w = w(\cdot, p)$ is either constant $w \equiv c$ or there are infinitely many discrete solutions to the equation $w(r) = c$, which interlace with the critical points of w.

Proof. We define the angular variable

$$\varphi := \arctan \frac{r^{n-1} w_r}{w - c}$$

$$= \operatorname{arccot} \frac{w - c}{r^{n-1} w_r}.$$

This is well defined outside of

$$M := \{r \geqq 0 \mid w(r, p) = c,\ w_r(r, p) = 0\},$$

which is both closed and relatively open in $\mathbb{R}_0^+$. The latter follows from the local uniqueness for solutions of (9). Thus M either coincides with $\mathbb{R}_0^+$, which induces $w \equiv c$, or is empty.

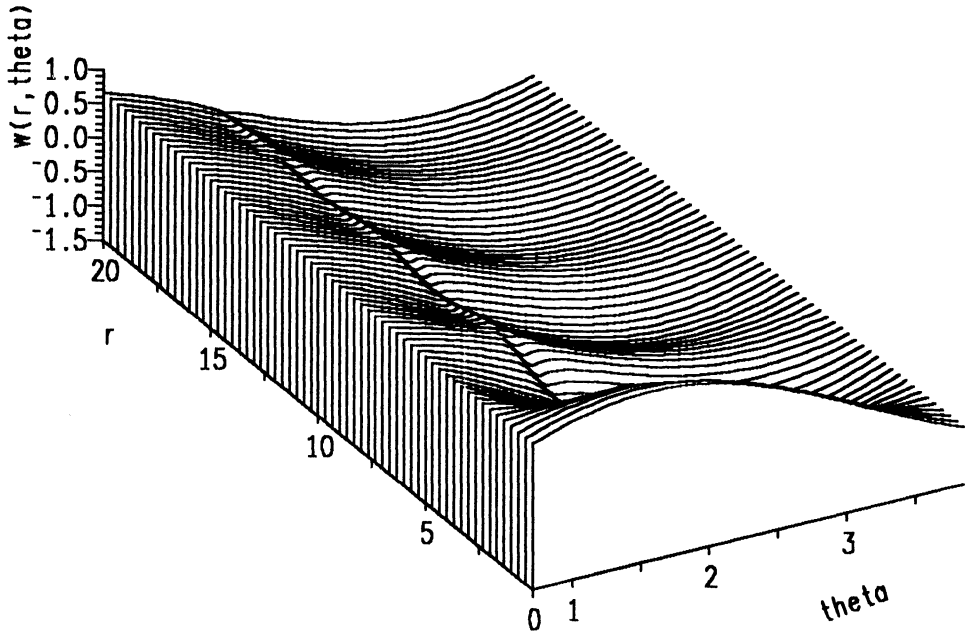


FIG. 4. $w(r, \vartheta)$ *for* $0.71046 = \vartheta_0 < \vartheta < \vartheta_0 + \pi$ *and* $0 < r < 20.0$. *We used the nonlinearity* $f(u) = u - \operatorname{arccot} u$ *and* $n = 3$.

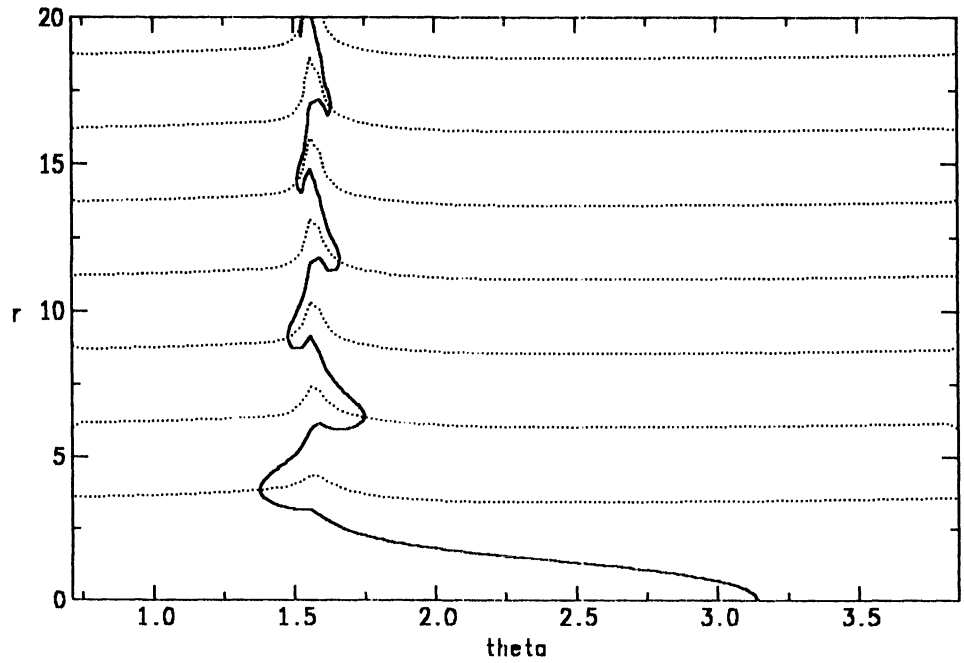


FIG. 5. *The zeros of* $w(r, \vartheta)$ *(solid line) and of* $w_r(r, \vartheta)$ *(dotted lines).*

In the latter case we take the derivative of $\tan \varphi(r)$ and get

$$\frac{\varphi'(r)}{\cos^2 \varphi(r)} = \frac{(r^{n-1}w_r)_r(w-c) - r^{n-1}(w_r)^2}{(w-c)^2}$$

$$= -r^{n-1}\frac{g(w)}{w-c} - r^{-n+1}\tan^2 \varphi(r),$$

and thus

$$\varphi'(r) = -r^{n-1}\frac{g(w)}{w-c}\cos^2 \varphi(r) - r^{-n+1}\sin^2 \varphi(r) < 0.$$

The last inequality holds because the expression $g(w)/(w-c)$ exceeds a positive constant $\varepsilon > 0$, which in turn follows from the admissibility of g via

$$\lim_{w\to\pm\infty}\frac{g(w)}{w-c} = \lim_{w\to\pm\infty}\frac{g(w)}{w} = a > 0,$$

$$\lim_{w\to c}\frac{g(w)}{w-c} = g_w(c) > 0.$$

This already implies the interlacing of the solutions of $w(r) = c$ with the critical points of w, because $\varphi(r)$ being strictly decreasing hits the negative multiples of $\pi/2$ in decreasing order. An even multiple corresponds to a zero of w_r, an odd multiple to a solution of $w(r) = c$.

We still must show that there are infinitely many solutions to the equation $w(r) = c$. We do this by comparison with the Bessel function J_ν of order $\nu = (n-2)/2$, which is known to have infinitely many zeros.

Suppose that $w(r)-c$ does not change sign between two consecutive zeros a, b of the solution $v(r) := r^{-\nu} J_\nu(r\sqrt{\varepsilon})$ of

$$v'' + \frac{n-1}{r} v' + \varepsilon \cdot v = 0.$$

Replacing v by $-v$, if necessary, we can assume that $v(w-c) \geqq 0$ in $]a, b[$. Then we have

$$\begin{aligned} 0 &< \int_a^b \left(\frac{g(w)}{w-c} - \varepsilon \right) \cdot v \cdot (w-c) r^{n-1}\, dr \\ &= \int_a^b \left[v''(w-c) - w''v + \frac{n-1}{r} (v'(w-c) - w'v) \right] r^{n-1}\, dr \\ &= \int_a^b [r^{n-1}(v'(w-c) - w'v)]_r\, dr \\ &= b^{n-1} v'(b)(w(b)-c)) - a^{n-1} v'(a)(w(a)-c)) \\ &\leqq 0. \end{aligned}$$

This obviously is a contradiction and $w-c$ must change sign in $]a, b[$. □

COROLLARY 2.4. *Either both* $w(r, \vartheta) \equiv \sin \vartheta_0$ *and* $w_r(r, \vartheta) \equiv 0$ *are constant, which forces* $\vartheta = \vartheta_0$, *or the zeros of* $w_r(\cdot, \vartheta)$ *and* $w(\cdot, \vartheta) - \tan \vartheta_0 \cdot \cos \vartheta$ *are discrete, simple, and interlace.*

Proof. Lemma 2.3 implies that $w(\cdot, \vartheta)$ either equals or oscillates around the value $c := \tan \vartheta_0 \cdot \cos \vartheta$. For constant w the initial condition leads to $c = w(0, \vartheta) = \sin \vartheta$, which implies $\vartheta = \vartheta_0$. In the other case the zeros of w_r and $w-c$ are discrete and interlace. This already proves the simplicity of the zeros of $w-c$. As another consequence, w_r and $f(w/\cos \vartheta)$ cannot have common zeros. For any zero of w_r the differential equation therefore leads to

$$w_{rr} = -\cos \vartheta f\left(\frac{w}{\cos \vartheta} \right) \neq 0,$$

which excludes common zeros of w_r and w_{rr}. Thus the zeros of w_r are simple. □

COROLLARY 2.5. *Suppose,* (r, ϑ) *is a common solution of* $w(r, \vartheta) = 0$ *and* $w_r(r, \vartheta) = 0$. *If* r *is the* k*th positive zero of* $w_r(\cdot, \vartheta)$ *and if* $\vartheta_0 < \vartheta < \vartheta_0 + \pi$, *then* ϑ *satisfies*

$$\text{sign}\,(\cos \vartheta \cdot f(0)) = (-1)^k. \tag{10}$$

Proof. According to Corollary 2.4 $w(\cdot, \vartheta)$ oscillates around the value $c := \tan \vartheta_0 \cdot \cos \vartheta$, which must be nonzero to allow for common zeros of w and w_r. This excludes $\vartheta_0 = 0$ and $\vartheta \equiv \pi/2 (\pi)$. Using the differential equation, we arrive at

$$-w_{rr}(r, \vartheta) = g(\cos \vartheta, 0) = \cos \vartheta \cdot f(0). \tag{11}$$

Due to the oscillation any two consecutive zeros of w_r differ in sign w_{rr}. Computing this sign for $r = 0$, we find

$$-w_{rr}(0, \vartheta) = g(\cos \vartheta, \sin \vartheta) = \cos \vartheta \cdot f(\tan \vartheta) > 0,$$

provided $\vartheta \in]\vartheta_0, \vartheta_0 + \pi[$ and $\vartheta \neq \pi/2$. For the kth positive zero we therefore get $-\text{sign}\, w_{rr}(r, \vartheta) = (-1)^k$. In conjunction with equation (11) this completes the proof. □

LEMMA 2.6. *Let $\vartheta_0 \neq 0$. Then every component $\mathscr{C}$ of zeros of $w(r, \vartheta)$ is contained in a strip $\mathbb{R} \times]\vartheta_0 + k\pi, \vartheta_0 + (k+1)\pi[$ for some $k \in \mathbb{Z}$.*

Proof. Let $k \in \mathbb{Z}$ and let

$$\mathscr{D}_k := \{(r, \vartheta) \in \mathscr{C} \mid \vartheta \leqq \vartheta_0 + k\pi \text{ or } \vartheta \geqq \vartheta_0 + (k+1)\pi\}.$$

This set is closed by construction. It is also open in $\mathscr{C}$, because $(r, \vartheta_0 + k\pi)$, $(r, \vartheta_0 + (k+1)\pi) \notin \mathscr{C} \supset \mathscr{D}_k$. Thus it is either empty or coincides with $\mathscr{C}$. The latter must be true for at least one $k \in \mathbb{Z}$. □

2.3. Arriving at the proof.

Proof of 2.1. According to Corollary 2.4 we can apply the implicit function theorem to obtain parametrizations $R_k(\vartheta)$ for the zeros of $w_r(r, \vartheta)$, provided $\vartheta \not\equiv \vartheta_0(\pi)$. For the remaining values of ϑ the branches described by R_k intersect with the trivial solution $\{(r, \vartheta_0 + m\pi) \mid r \geqq 0, m \in \mathbb{Z}\}$. To deal with these bifurcations we apply the above argument to the function

$$v(r, \vartheta) := \begin{cases} \dfrac{u(r, \tan \vartheta) - \tan \vartheta_0}{\tan \vartheta - \tan \vartheta_0} & \text{if } \vartheta \not\equiv \vartheta_0(\pi), \\ u_p(r, \tan \vartheta_0) & \text{if } \vartheta \equiv \vartheta_0(\pi). \end{cases}$$

Note that u_p solves the differential equation

$$u_p'' + \frac{n-1}{r} u_p' + f'(\tan \vartheta_0) u_p = 0 \quad \text{in } \mathbb{R}^+,$$

where $\hat{f}(u) := f'(\tan \vartheta_0)u$ is an admissible function. Thus Corollary 2.4 implies that u_p has infinitely many simple zeros and the implicit function theorem yields the desired parametrization over ϑ.

Because of the uniqueness of the solutions of the initial value problem (8) we have $w(r, \vartheta + \pi) = -w(r, \vartheta)$. This implies the periodicity of the zero set of w_r and thus the periodicity of the functions $R_k(\vartheta)$. □

LEMMA 2.7. *Let $\vartheta_0 \neq 0$ hold. Then there is a continuum $\mathscr{C}_0$ of zeros of $w(r, \vartheta)$ satisfying the following:*

- *$\mathscr{C}_0$ lies in the strip $\mathbb{R}_0^+ \times]\vartheta_0, \vartheta_0 + \pi[$.*
- *$\mathscr{C}_0$ contains $(0, \pi)$ or $(0, 0)$, depending on the sign of ϑ_0.*
- *$\mathscr{C}_0$ is unbounded.*

Proof. Looking first for the zeros satisfying $r = 0$ we get $0 = w(0, \vartheta) = \sin \vartheta$ and thus $\vartheta = k\pi$ for some $k \in \mathbb{Z}$. Since the zero should be contained in the interval $]\vartheta_0, \vartheta_0 + \pi[$ this reduces to $\vartheta = 0$ or $\vartheta = \pi$ depending on the sign of ϑ_0.

We extend w continuously for negative r using $w(r, \vartheta) = w(-r, \vartheta)$ and apply Theorem 1.1 to $F = w$, $G = H = \{1\}$, and $\lambda = r$. In this context the above zero is H-regular and the zeros nearby lie on a curve $\vartheta(r)$, $|r| < \varepsilon$. Now we choose $a = 0$, $b = 2\varepsilon$, and $\mathscr{S}_1$ to be the part of the curve satisfying $r \geqq 0$. By construction no element of $\mathscr{S}_1$ lies above b, but one point of $\mathscr{S}_1$ lies above a. This point is a special boundary point having index $i_H(0, k\pi) = \pm 1$. Theorem 1.1 now establishes the existence of some continuum $\mathscr{C}_0$ intersecting with the curve $\vartheta(r)$ in some point with $r > 0$.

This continuum must be unbounded by Theorem 1.1, if we exclude a second intersection of $\mathscr{S}_1$ with $\mathscr{C}_0$. To do this, note that the set

$$\mathscr{D} := \{(r, \vartheta) \in \mathscr{C}_0 \mid r \leqq 0\},$$

is both closed and open in $\mathscr{C}_0$, the first by construction. To see the openness, let $(0, \vartheta) \in \mathscr{D}$. By Lemma 2.6 this point must coincide with the special boundary point of

$\mathscr{S}_1$. According to Theorem 1.1 all zeros of w in some neighbourhood of this point are either in $\mathscr{S}_1$ or in $\mathscr{C}_0$. The latter must satisfy $r \leqq 0$ and are thus contained in $\mathscr{D}$.

Since $\mathscr{C}_0 \neq \mathscr{D}$, we have $\mathscr{D} = \varnothing$ because $\mathscr{C}_0$ is connected. This excludes a second intersection and this completes the proof. □

Proof of Theorem 2.2.

Case $\vartheta_0 \neq 0$. We replace the continuum $\mathscr{C}_0$ found in the previous lemma by its topological component, which we also denote by $\mathscr{C}_0$. Then the first two requirements of Theorem 2.2 are already satisfied, the latter by Lemma 2.6. It mainly remains to prove that this component intersects with the graphs of the functions R_k and crosses the line $\vartheta = \pi/2$ between any two consecutive of those intersections. If the first were not the case the set

$$\mathscr{D} := \{(r, \vartheta) \in \mathscr{C}_0 | r \leqq R_k(\vartheta)\}$$

would be both open and closed in $\mathscr{C}_0$. If the line $\vartheta = \pi/2$ were not crossed this would hold for the set

$$\mathscr{D} := \{(r, \vartheta) \in \mathscr{C}_0 | r \leqq R_k(\vartheta)\}$$
$$\cup \{(r, \vartheta) \in \mathscr{C}_0 | r \leqq R_{k+1}(\vartheta) \text{ and } \operatorname{sign}(\cos \vartheta \cdot f(0)) = (-1)^k\}.$$

In both cases a contradiction to the connectedness of $\mathscr{C}_0$ would result. Note that in the second argument we already used the crossing condition as established in Corollary 2.5. Finally, to pinpoint the places where $\mathscr{C}_0$ crosses the line $\vartheta = \pi/2$ we note that $w(r, \pi/2)$ is essentially a Bessel function. More precisely, $J(r\sqrt{a}) := r^\nu w(r, \pi/2)$ with $\nu := (n-2)/2$ is a bounded solution of the differential equation

$$J'' + \frac{1}{r} J' + \left(1 - \frac{\nu^2}{r^2}\right) J = 0 \quad \text{in } \mathbb{R}^+$$

and must coincide with the Bessel function J_ν (up to some constant factor).

Case $\vartheta_0 = 0$. As with the proof of Theorem 2.1 we get the functions $r_k(\vartheta)$ by application of the implicit function theorem to w, if $\vartheta \neq 0$, or to

$$v(r, \vartheta) := \begin{cases} \dfrac{u(r, \tan \vartheta)}{\tan \vartheta} & \text{if } \tan \vartheta \neq 0, \\ u_p(r, 0) & \text{if } \tan \vartheta = 0 \end{cases}$$

in the other case. The zeros of w and $v(\cdot, 0)$ are simple by Corollary 2.4. Because of the uniqueness of the solutions of the initial value problem (8) we have $w(r, \vartheta + \pi) = -w(r, \vartheta)$. This implies the periodicity of the zero set of w and thus the periodicity of the functions $r_k(\vartheta)$. □

3. Bifurcation by remote control.

3.1. Designing the remote control. Because of Theorem 2.2 there is a component $\mathscr{C}_0$ of radially symmetric solutions (ϑ, R, u) of the boundary value problem

$$\begin{aligned} &\Delta u + f(u) = 0 \quad \text{in } \mathbb{B}_R(0), \\ &u = 0 \quad \text{in } \partial \mathbb{B}_R(0), \\ &u(0) = \tan \vartheta, \end{aligned} \tag{12}$$

which follows a slalom course around the "flags" $(R_k(\pi/2), \pi/2)$. The functions $R_k(\vartheta)$ describe the solutions of the corresponding Neumann problem (Theorem 2.1). The value $\vartheta = \pi/2$ was introduced by rescaling. Naively, solutions $(\pi/2, R, u)$ are solutions

of problem (12) satisfying $u(0)=\infty$. Surprisingly, these solutions provide information about bifurcations with symmetry breaking taking place somewhere else.

THEOREM 3.1. *For all admissible f satisfying $f(0)\neq 0$ the following hold:*

(1) *If the line $\vartheta=\pi/2$ is taken away, the slalom course $\mathscr{C}_0$ established in Theorem 2.2 disconnects into countably many components. Except for the component that hits the line $r=0$, each of these components connect two consecutive elements $(r_k, \pi/2)$, $(r_{k+1}, \pi/2)$ in the intersection of $\mathscr{C}_0$ with the line $\vartheta=\pi/2$. This pair of points uniquely determines the component, the closure (with respect to $\mathbb{R}^2$) of which we denote by $\mathscr{C}_k$. The values $r_i\sqrt{a}=[J_\nu]_i$ are given by the zeros of the Bessel function J_ν of order $\nu=(n-2)/2$.*

(2) *Suppose that the set*

$$N_k := \{[J_\nu]_k < R < [J_\nu]_{k+1} | J_{\nu+k}(R)=0 \text{ for } 0\leqq\kappa\in\mathbb{Z}\}$$

has odd cardinality. Then there is a continuum $\mathscr{C}$ of solutions (ϑ, R, u) of problem (12) having minimum symmetry $O(n-1)$ and satisfying the following.

(a) *$\mathscr{C}$ starts in $\mathscr{C}_k$.*

(b) *$\mathscr{C}$ contains only points (ϑ, R, u) with $R>0$ and $\vartheta\not\equiv\pi/2(\pi)$, i.e., only "true" solutions of problem (12).*

(c) *$\mathscr{C}$ does not contain any interior point of*

$$\mathscr{S}_0 := \{(\vartheta, R, u) | (\vartheta+m\pi, R, u)\in\mathscr{C}_0 \text{ for some } m\in\mathbb{Z}\}.$$

(d) *$\mathscr{C}$ is unbounded in R or $\|u\|_{C^2}$ or intersects with $\mathscr{S}_0$ outside $\mathscr{C}_k$.*

(3) *Now suppose that the set*

$$N_k^* := \{[J_\nu]_k < R < [J_\nu]_{k+1} | J_{\nu+\kappa}(R)=0 \text{ for } 0\leqq\kappa\equiv 0(2)\}$$

has odd cardinality. Then there is a continuum $\mathscr{C}$ of solutions (ϑ, R, u) of problem (12) having minimum symmetry $O(n-1)\times\mathbb{Z}_2$ and satisfying (a)-(d) *in* (2) *above.*

The cardinalities of the sets N_k and N_k^* do not depend on the nonlinearity f and can be easily computed. A few of them are listed in the Appendix. Note that in almost all cases at least one of the cardinalities N_k and N_k^* is odd, thereby guaranteeing bifurcation with symmetry breaking. In particular, N_1^* is odd for every space dimension that is listed in the Appendix. Thus, due to the above theorem, a continuum $\mathscr{C}$ of solutions (ϑ, R, u) of problem (12) having minimum symmetry $O(n-1)\times\mathbb{Z}_2$ bifurcates somewhere from $\mathscr{C}_1$.

Recall that Smoller and Wasserman [14] already found a bifurcation from $\mathscr{C}_1$, which was later proven to be global by Cerami [1]. But the local analysis shows that solutions sufficiently near to Smoller's bifurcation point either belong to $\mathscr{C}_1$ or have a symmetry conjugate to $O(n-1)$. Since the above branch has minimum symmetry $O(n-1)\times\mathbb{Z}_2$, we have found a bifurcation different from that of Smoller. Smoller also proved (his main argument is contained in Lemma 3.7 below) that his bifurcation point is the only one for positive solutions. Thus ours cannot be a positive solution of problem (12). So it has to be Mexican hat shaped, because this is the only other type of solution living on $\mathscr{C}_1$.

To establish their results, both Smoller and Cerami use a transversality condition. We will now prove that is automatically satisfied and can be dropped. We will reformulate their result as a theorem in our setting—which then is a consequence of Corollary 1.2—but to save space we drop the complete description of the situation near the bifurcation point—which needs local arguments such as the Lyapunov-Schmidt reduction.

THEOREM 3.2. *Let f be twice continuously differentiable and admissible with $f(0)\neq 0$. Let (R_1, ϑ_1) denote an intersection of the slalom course $\mathscr{C}_0$ established in Theorem 2.2 with the graph of the function $R_1(\vartheta)$. Besides we assume $w_\vartheta(R_1, \vartheta_1)\neq 0$.*

Then there is a continuum $\mathscr{C}$ *of solutions* (ϑ, R, u) *of problem* (12) *having minimum symmetry* $O(n-1)$ *and satisfying the following.*

(1) $\mathscr{C}$ *starts in* (R_1, ϑ_1).

(2) $\mathscr{C}$ *contains only points* (ϑ, R, u) *with* $R>0$ *and* $\vartheta \not\equiv \pi/2(\pi)$, *i.e., only "true" solutions of problem* (12).

(3) $\mathscr{C}$ *does not contain any interior point of*

$$\mathscr{S}_0 := \{(\vartheta, R, u) | (\vartheta + m\pi, R, u) \in \mathscr{C}_0 \text{ for some } m \in \mathbb{Z}\}.$$

(4) $\mathscr{C}$ *is unbounded in* R *or* $\|u\|_{C^2}$ *or intersects with* $\mathscr{S}_0$ *in a point different from* (R_1, ϑ_1).

Note that having a solution of problem (12) with small symmetry we get a whole bunch of them by considering conjugate symmetries. In detail, let $\mathscr{C}$ be the bifurcating continuum found in Theorem 3.1 and let $H \supseteq O(n-1)$ be the prescribed minimum symmetry. Then $[\gamma]\mathscr{C} = \gamma\mathscr{C}$ is also a continuum of solutions of problem (12) for every left coset $[\gamma] \in O(n)/H$. $[\gamma]\mathscr{C}$ is unbounded if and only if $\mathscr{C}$ is. $[\gamma]\mathscr{C}$ and $\mathscr{C}$ intersect with $\mathscr{S}_0$ in identical points. The symmetry of some point $\gamma(\vartheta, R, u) \in [\gamma]\mathscr{C}$ is conjugate to the symmetry of $(\vartheta, R, u) \in \mathscr{C}$.

3.2. Some technical details. Given the function spaces

$$X := W^{2,p}(\mathbb{B}_1(0)) \cap W_0^{1,p}(\mathbb{B}_1(0)),$$

$$Y := L_p(\mathbb{B}_1(0)),$$

where $2 - n/p > 1$, we define $F: \mathbb{R}^2 \times X \mapsto \mathbb{R} \times Y$ in the following way:

$$F(\vartheta, R, u)(x) := (u(0) - \sin\vartheta, \Delta u(x) + R^2 g(\cos\vartheta, u(x))),$$

where

$$g(q, w) = \begin{cases} qf(w/q) & \text{if } q \neq 0, \\ aw & \text{if } q = 0, \end{cases}$$

and f is assumed to be admissible. F is equivariant with respect to the group $O(n)$ operating on X and Y, respectively, by

$$[\gamma u](x) := u(\gamma^{-1}x) \quad \text{for all } \gamma \in O(n), \quad u \in Y.$$

The spaces X and Y are chosen to satisfy the following:

(1) Both $\Delta: X \mapsto Y$ and, consequently,

$$B := \begin{pmatrix} 1 & 0 \\ 0 & \Delta \end{pmatrix} : \mathbb{R} \times X \mapsto \mathbb{R} \times Y$$

have a bounded inverse (Simader [12, Thm. 10.10, p. 184] or Zeidler [19, Thm. 6.8a, p. 259]).

(2) The map $u \mapsto u(0)$ from X to $\mathbb{R}$ is continuous.

(3) The map $G: \mathbb{R} \times C^0(\overline{\mathbb{B}_1(0)}) \mapsto Y$ defined via $G(q, u)(x) := g(q, u(x))$ is continuous. If $f \in C^k(\mathbb{R})$, then G is k times continuously (Fréchet-) differentiable for $q \neq 0$.

(4) Both the embedding $\iota: X \hookrightarrow C^0(\overline{\mathbb{B}_1(0)})$ and, consequently, the map

$$F(\vartheta, R, u) - B(R, u) = (u(0) - R - \sin\vartheta, R^2 G(\cos\vartheta, \iota u))$$

are compact.

The function spaces and in particular the condition $2 - n/p > 1$ are chosen to make the proof of the following lemma work. This lemma is the key ingredient to the proof of the continuous dependence of the partial derivative $D_{R,u}F(\vartheta, R, u)$ on (ϑ, R, u) in

the neighbourhood of some zero $(\pi/2, R_0, u_0)$ of F. This being granted we are able to apply the implicit function theorem.

LEMMA 3.3. *Let* $2-n/p>1$ *and let zero be a regular value of*

$$u_0 \in X := W^{2,p}(\mathbb{B}_1(0)) \cap W_0^{1,p}(\mathbb{B}_1(0) \overset{\iota}{\hookrightarrow} C^1(\overline{\mathbb{B}_1(0)}).$$

Then $D_u G(q, \iota u)$ *is continuous in a neighbourhood of* $(0, u_0)$.

Proof. Let (q_i, u_i) be a sequence converging to $(0, u)$ in $\mathbb{R} \times X$. Then we have the following:

$$\|G_u(q_i, \iota u_i)w - G_u(0, \iota u)w\|_{L_p}^p = \int_{\mathbb{B}(0)} |g_u(q_i, \iota u_i(x))w(x) - a \cdot w(x)|^p \, dx$$

$$\leqq \int_{\mathbb{B}(0)} \left| f'\left(\frac{\iota u_i(x)}{q_i}\right) - a \right|^p dx \|w\|_{C^0}^p.$$

The sequence ιu_i converges to ιu pointwise in C^0. So the expression under the integral tends to zero for all x with $\iota u(x) \neq 0$. By the Lebesgue dominated convergence theorem the integral and so the operator norm of $G_u(q_i, \iota u_i) - G_u(0, \iota u) \to 0$—provided the zeros of ιu have measure zero.

If $\|u - u_0\|_{W^{2,p}}$ is sufficiently small, so is the distance $\|\iota u - \iota u_0\|_{C^1} \leqq C\|u - u_0\|_{W^{2,p}}$ with respect to the C^1-norm. Thus zero is not only a regular value of ιu_0, but also of ιu. Then the zero set of ιu is a manifold of dimension $n-1$ having (n-dimensional) Lebesgue measure 0. □

LEMMA 3.4. *The eigenvalues of the Laplace operator on the sphere* S^{n-1} *are* $-\kappa(\kappa + n-2)$, *where* $0 \leqq \kappa \in \mathbb{Z}$. *In every eigenspace there is (up to constant multiples) precisely one eigenfunction having minimum symmetry* $O(n-1)$. *This eigenfunction is of the form* $P_\kappa(x_n/\|x\|)$, *where* P_κ *denotes a polynomial of degree* κ. *This polynomial is an even function if* κ *is even, and an odd function otherwise.*

Proof. Let Φ_κ be an eigenfunction of the spherical Laplacian for the eigenvalue λ_κ. Furthermore, let $\xi_0 \in S^{n-1}$ be a point on the sphere satisfying $\Phi_\kappa(\xi_0) \neq 0$ and let $\gamma O(n-1)\gamma^{-1}$ be the symmetry of ξ_0. Then

$$\tilde{\Phi}_\kappa(\xi) := \int_{O(n-1)} \sigma\gamma^{-1}\Phi_\kappa(\xi) \, d\sigma = \int_{O(n-1)} \Phi_\kappa(\gamma\sigma^{-1}\xi) \, d\sigma$$

is an $O(n-1)$ invariant eigenvector for the eigenvalue λ_κ. This eigenvector does not vanish because $\tilde{\Phi}_\kappa(\gamma^{-1}\xi_0) = \Phi_\kappa(\xi_0) \neq 0$.

For $n>2$ the eigenfunction $\tilde{\Phi}_\kappa$ being invariant under $O(n-1)$ only depends on the "latitude" $\varphi := \arccos(x_n/\|x\|) \in [0, \pi]$ of the sphere S^{n-1}. Making the ansatz $\tilde{\Phi}_\kappa(\varphi) = P_\kappa(\cos\varphi)$, where P_κ is a polynomial with degree $P_\kappa = \kappa$, for $x \in [-1, 1]$ this polynomial is a bounded solution of the equation

$$(1-x^2)P_\kappa''(x) - (n-1)xP_\kappa'(x) - \lambda_\kappa P_\kappa(x) = 0.$$

We can compute the coefficients α_ν of P_κ recursively using the formula

$$(\lambda_\kappa + \nu(\nu+n-2))\alpha_\nu = \begin{cases} 0 & \text{if } \nu = \kappa \text{ or } \nu = \kappa - 1, \\ (\nu+1)(\nu+2)\alpha_{\nu+2} & \text{if } 0 \leqq \nu < \kappa - 1. \end{cases}$$

We immediately get $\lambda_\kappa = -\kappa(\kappa+n-2)$ and $\alpha_\nu = 0$ for all $\nu \not\equiv \kappa(2)$. Thus the polynomial either is an even or odd function and there is—up to a constant multiple—only one such polynomial P_κ.

The functions $\tilde{\Phi}_\kappa(\varphi) = P_\kappa(\cos\varphi)$ are pairwise orthogonal with respect to the scalar product

$$\langle u, v\rangle := \int_0^\pi u(\varphi)\overline{v(\varphi)}\sin^{n-2}(\varphi)\,d\varphi$$

and constitute a basis of $L_2([0,\pi])$ (with the above scalar product) due to the Weierstrass approximation theorem. This way we get all eigenfunctions that are invariant under $O(n-1)$.

The statements of the lemma, in particular $\tilde{\Phi}_\kappa(\varphi) = P_\kappa(\cos\varphi)$, also hold for $n=2$, but the proof must be different. As above, we choose $\varphi := \arccos(x_2/\|x\|) \in [0, 2\pi]$, but now we have an explicit formula

$$\tilde{\Phi}_\kappa(\varphi) = \mathrm{Re}(c\,e^{i\kappa\varphi}) \quad \text{where } c\in\mathbb{C}, \quad \kappa\in\mathbb{Z}.$$

Because of the invariance under $O(1)$ we have $\tilde{\Phi}(\varphi) = \tilde{\Phi}(-\varphi)$, which induces $\mathrm{Im}\,c = 0$. Thus $\tilde{\Phi}(\varphi)$ is a homogeneous polynomial of degree κ in $\cos\varphi$ and $\sin\varphi$. There are only even powers of $\sin\varphi$, and we can replace $\sin^2\varphi$ by $1-\cos^2\varphi$. We obtain a polynomial $P_\kappa(\cos\varphi)$ as desired. □

THEOREM 3.5. (1) *$(\pi/2, R, u)$ is a zero of F if and only if u is an eigenvector of $-\Delta$ for the eigenvalue aR^2 satisfying $u(0)=1$. All these solutions are radially symmetric and $R\sqrt{a}$ is a zero of the Bessel function J_ν, where $\nu = (n-2)/2$.*

(2) *All solutions sufficiently near to one of the above are radially symmetric and lie on a continuous curve parametrized by ϑ.*

(3) *Let $H = O(n-1)$. Then we have $i_H(\pi/2, R, u) = (-1)^\beta$, where*

$$\beta = \#\{0 < r < R\sqrt{a} \mid J_{\nu+\kappa}(r) = 0 \text{ for some } 0 \leqq \kappa \in \mathbb{Z}\}.$$

(4) *Let $H = O(n-1)\times\mathbb{Z}_2$. Then we have $i_H(\pi/2, R, u) = (-1)^\beta$, where*

$$\beta = \#\{0 < r < R\sqrt{\alpha} \mid J_{\nu+\kappa}(r) = 0 \text{ for some } 0 \leqq \kappa \equiv 0(2)\}.$$

Proof. (1) For any zero (ϑ, R, u) of F with $\vartheta = \pi/2$ the function u by construction is an eigenfunction of $-\Delta$ for the eigenvalue $\lambda = \alpha R^2$ satisfying $u(0)=1$. Passing to polar coordinates $(r,\xi)\in\mathbb{R}\times S^{n-1}$, we can expand u in a series

$$u(r,\xi) = \sum_{\kappa=0}^{\infty} a_\kappa(r)\Phi_\kappa(\xi),$$

where Φ_κ denotes an eigenfunction of the spherical Laplacian for the eigenvalue $-\kappa(\kappa+n-2)$ and $J_{\nu+\kappa}(r\sqrt{\lambda}) := r^\nu a_\kappa(r)$ is a bounded solution of the Bessel differential equation and therefore must coincide (up to constant multiple) with the Bessel function of order $\nu+\kappa$. Since u satisfies Dirichlet boundary conditions, we obtain $J_{\nu+\kappa}(\sqrt{\lambda}) = J_{\nu+\kappa}(R\sqrt{\alpha}) = 0$. For arbitrary integers $\kappa, \mu \geqq 0$, $J_{\nu+\kappa}$ and $J_{\nu+\mu}$ cannot have common zeros.[5] This leads to

$$u(r,\xi) = a_\kappa(r)\Phi_\kappa(\xi)$$

for some $\kappa \geqq 0$. Since $u(0)=1$, this implies $a_\kappa(0)\neq 0$ and Φ_κ constant. Due to Lemma 3.4 this can only happen for $\kappa = 0$. Since Φ_κ is $O(n)$-invariant, so is u.

(2) $\nabla u_0(x) = a_0'(\|x\|)\cdot x/\|x\|$ does not vanish for any zero of $u_0(x) = a_0(\|x\|)$. Thus zero is a regular value of u_0, which implies the continuity of $D_{R,u}F(\vartheta, R, u)$ by

[5] See Watson [17, p. 485] for Siegel's proof of the conjecture of Bourget.

Lemma 3.3. In order to compute the kernel of $L = D_{R,u}F(\pi/2, R, u_0)$ we start with the equation

$$0 = L(S, v) = (v(0), \Delta v + aR^2 v + 2aRSu).$$

Using the L_2 scalar product $\langle , \rangle$ this leads to

$$\begin{aligned} -2aRS\langle u, u\rangle &= \langle \Delta v + aR^2 v, u\rangle \\ &= \langle v, \Delta u + aR^2 u\rangle = 0, \end{aligned}$$

which in turn implies $S = 0$ and $v(x) = \tilde{c}u_0(x)$. Because of $v(0) = 0$ we get $\tilde{c} = 0$. The kernel therefore is trivial and applying the implicit function theorem to F we can parametrize the zeros of F by ϑ. We can also apply the implicit function theorem to $F^{O(n)}$ and we get a second curve of zeros of F with symmetry $O(n)$. By the uniqueness part of the implicit function theorem these two curves must coincide wherever they both exist.

(3) and (4). To compute the spectrum of $B^{-1}L = B^{-1}D_{R,u}F(\pi/2, R, u)$ we expand an average solution $(S, v) \in \mathbb{R} \times X \subseteq \mathbb{R} \times L_2(\mathbb{B}_1(0))$ of the equation

$$\begin{aligned} 0 &= B(B^{-1}L - \mu)(S, v) \\ &= L(S, v) - \mu B(S, v) \\ &= (v(0) - \mu S, (1-\mu)\Delta v + aR^2 v + 2aRSu) \end{aligned}$$

in a series $v = \sum_{i \geq 0} \alpha_i u_i$ of eigenfunctions for eigenvalues λ_i of $-\Delta$. The numbering of the eigenvalues is chosen to satisfy $u_0 = u$, $\lambda_0 = aR^2$. Inserting this in the above equation we get

$$\begin{aligned} \alpha_i(-\lambda_i(1-\mu) + aR^2) &= 0 \quad \text{for } i > 0, \\ \alpha_0 aR^2\mu + 2aRS &= 0 \qquad \text{for } i = 0. \end{aligned}$$

At most one α_i with $i > 0$ can be nonzero. If this is the case, we obtain an eigenvalue

$$\mu_i = 1 - \frac{aR^2}{\lambda_i}$$

for the eigenvector

$$v_i = \alpha_i u_i + \alpha_0 u_0 = \alpha_i \left(u_i - \frac{2u_i(0)}{R\mu_i^2 + 2} u_0 \right).$$

For the computation of S and α_0 we have used the equations

$$\begin{aligned} \alpha_0 aR^2\mu_i + 2aRS &= 0, \\ \alpha_i u_i(0) + \alpha_0 &= v_i(0) = \mu_i S. \end{aligned}$$

If $\alpha_i = 0$ vanishes for all $i > 0$, the eigenvector has the form $v = \alpha_0 u_0$. S and μ can be obtained from the following equations:

$$\alpha_0 aR^2\mu + 2aRS = 0, \qquad \mu S = \alpha_0,$$

which leads to

$$\mu = \pm i\sqrt{2/R}.$$

To compute the indices we must determine the algebraic multiplicity of all negative real eigenvalues of $(B^{-1}L)^H$. We will show that all relevant eigenvalues are semisimple,

but treat only the case $H=\{1\}$: If there is no generalized eigenvector, it cannot have any prescribed symmetry. Let $\mu_j = 1 - aR^2/\lambda_j$ be an eigenvalue with generalized eigenvector $(B^{-1}L-\mu_j)^2(\tilde{S},\tilde{v})=0$. We again expand $\tilde{v}$ in a series $\tilde{v}=\sum_{i\geq 0}\tilde{\alpha}_i u_i$. Put this into

$$(L-B\mu_j)(\tilde{S},\tilde{v}) = B(S,v),$$
$$(L-B\mu_j)(S,v)=0,$$

and obtain

$$\tilde{\alpha}_i(-\lambda_i(1-\mu_j)+aR^2)=0 \quad \text{for } i>0 \text{ and } i\neq j,$$
$$\tilde{\alpha}_i(-\lambda_i(1-\mu_j)+aR^2)=-\lambda_j\alpha_j \quad \text{for } i=j,$$
$$\tilde{\alpha}_0 aR^2\mu_j + 2aR\tilde{S} = \alpha_j\left(\frac{2aR^2u_j(0)}{R\mu_j^2+2}\right) \quad \text{for } i=0.$$

Considering the equation for $i=j$ we immediately get $\alpha_j=0$. Then $(\tilde{S},\tilde{v})$ already is an eigenvector in the usual sense and the eigenvalue is semisimple.

Prescribing minimum symmetry $O(n-1)$ we find by part (a) of the proof and Lemma 3.4 that for every i there is precisely one κ such that

$$v_i(x) = \alpha_i P_\kappa(x_n/\|x\|)\left(a_\kappa(\|x\|) - \frac{2a_\kappa(0)}{R\mu_i^2+2}\,a_0(\|x\|)\right).$$

Here P_κ denotes a polynomial of degree κ, which is an even function for κ even and an odd function otherwise and $r^\nu a_\kappa(r) =: J_{\nu+\kappa}(r\sqrt{\lambda})$ up to constant multiple is the Bessel function of order $\nu+\kappa$.

In the case where $H=O(n-1)$, all eigenspaces are one-dimensional. To calculate i_H we therefore only have to count real negative eigenvalues $\mu_i = 1-aR^2/\lambda_i$. This number just coincides with the number of zeros $0<\sqrt{\lambda_i}<R\sqrt{a}$ of the Bessel function $J_{\nu+\kappa}$. This proves part (c).

$O(n-1)\times\mathbb{Z}_2$ also contains the reflection at the plane $x_n=0$. Thus only eigenfunctions with P_κ even are invariant. This implies κ even—and the proof of part (d) is complete. □

3.3. Assembly of the proof parts.

Remark. Let $(\vartheta, R, u)\in\mathbb{R}^2\times X$ be a zero of F. Then u is contained in the set

$$u\in\tilde{X} := \{u\in C^{2,\alpha}(\overline{\mathbb{B}_1(0)}) \mid u\equiv 0 \text{ on } \partial\mathbb{B}_1(0)\}.$$

If moreover $R>0$ and $\vartheta\not\equiv\pi/2(\pi)$, then $\tilde{u}(x) := (\cos\vartheta)^{-1}u(x/R)$ is a classical solution of the Dirichlet problem (12).

Remark. To reduce the proof of Theorem 3.1 to an application of Theorem 1.1 we must identify the zeros of $w(r,\vartheta)$ with certain zeros of F. To do this we define the continuous map

$$\mathcal{W}: \mathbb{R}^2\mapsto\mathbb{R}^2\times W^{2,p}(\mathbb{B}_1(0)) \text{ via } \mathcal{W}(R,\vartheta) := (\vartheta, R, \mathcal{W}_0(R,\vartheta)),$$

where

$$\mathcal{W}_0(R,\vartheta)(x) := w(R\|x\|,\vartheta).$$

LEMMA 3.6. *$\mathcal{W}$ maps the zeros of w homeomorphically onto the zeros of $F^{O(n)}$. A zero (R,ϑ) of w is regular if and only if $\mathcal{W}(R,\vartheta)$ is $O(n)$-regular (in the sense of* § 1*). More precisely, if $\vartheta\not\equiv\pi/2$, every element in the kernel of $DF^{O(n)}(\mathcal{W}(R,\vartheta))$ is a linear combination of*

$$\mathcal{W}_\vartheta(R,\vartheta) \quad \text{and} \quad \mathcal{W}_R(R,\vartheta).$$

Proof. We only prove the second claim and restrict to the case $\vartheta \neq \pi/2$, since the other case was treated by Theorem 3.5. If (ρ, S, v) is an arbitrary element in the kernel, then

$$\tilde{v}(t) := v(t) - \rho w_\vartheta(R \cdot t, \vartheta) + S t w_r(R \cdot t, \vartheta)$$

solves the initial value problem

$$\tilde{v}'' + \frac{n-1}{t}\tilde{v}' + R^2 g_u(\cos\vartheta, w(R \cdot t, \vartheta))\tilde{v} = 0,$$

$$\tilde{v}(0) = \tilde{v}'(0) = 0.$$

By uniqueness [13, p. 157] we get $\tilde{v} \equiv 0$, which induces

$$(\rho, S, v) = \rho \mathcal{W}_\vartheta(R, \vartheta) + S\mathcal{W}_R(R, \vartheta).$$

Because of the Dirichlet boundary condition, ρ and S must satisfy the following relation:

$$0 = v(1) = \rho w_\vartheta(R, \vartheta) + S w_r(R, \vartheta).$$

Thus dim Kern $F^{O(n)}(\mathcal{W}(R, \vartheta)) = \dim \operatorname{Kern} \nabla w(r, \vartheta) \geqq 1$. □

Proof of Theorem 3.1(1). For $k \in \mathbb{Z}$ we consider the cells

$$\{(R, \vartheta) | R_{k-1}(\vartheta) < R < R_{k+1}(\vartheta) \text{ and } \operatorname{sign}(\cos\vartheta \cdot f(0)) = (-1)^k\},$$

where $R_{-k} := -R_k$ and $R_0 \equiv 0$. Let $\mathcal{C}_k$ be the closure of the intersection of the branch $\mathcal{C}_0$ as in Theorem 2.2 with cell $k > 0$. $\mathcal{C}_0$ intersects the boundary of cell k only in two points $(r_k, \pi/2)$, $(r_{k+1}, \pi/2)$, because the zeros $r_i/\sqrt{a}$ and the critical points $R_i/\sqrt{a}$ of the Bessel function J_ν interlace and every zero of F with $\vartheta = \pi/2$ is—by Theorem 3.5(1)—related to a zero of J_ν in the way described. Next we define the "projection" onto cell k to be

$$\Xi_k(r, \vartheta) := \begin{cases} (r, \vartheta) \in \mathcal{C}_0 & \text{if this point lies in cell } k, \\ (r_k, \pi/2) & \text{if not and } r < R_k(\vartheta), \\ (r_{k+1}, \pi/2) & \text{if not and } r > R_k(\vartheta). \end{cases}$$

By Theorem 3.5(2) all these projections are continuous maps on $\mathcal{C}_0$. Their images $\mathcal{C}_k$ therefore are connected and contain the two boundary points as desired. Finally, $\mathcal{C}_k \setminus \{(r_k, \pi/2), (r_{k+1}, \pi/2)\}$ is contained in some component of $\mathcal{C}_0 \setminus \mathbb{R} \times \{\pi/2\}$. Being the intersection of $\mathcal{C}_0$ with the open cell k, it is both open and closed in this component and thus has to coincide with it. This way we get all components, because the closures of all cells cover the plane $\mathbb{R}^2$.

(2) and (3). Choose $\mathcal{S}_1 := \mathcal{W}(\mathcal{C}_k)$, $\mathcal{S}_0$ as in Theorem 3.1 and

$$[a, b] := \begin{cases} [\vartheta_0, \pi/2] & \text{if } \operatorname{sign}(\vartheta_0) = (-1)^{k+1} \\ [\pi/2, \vartheta_0 + \pi] & \text{if } \operatorname{sign}(\vartheta_0) = (-1)^k \end{cases}$$

and apply Theorem 1.1. The assumptions are satisfied:

(1) $\mathcal{S}_1$ is compact, being the image of the compact set $\mathcal{C}_k$ under $\mathcal{W}$. $\mathcal{C}_k$ is compact, because it is closed and contained in a bounded cell.

(2) Only the two zeros $\mathcal{W}(r_k, \pi/2)$ and $\mathcal{W}(r_{k+1}, \pi/2)$ lie above the interval boundaries. Due to Theorem 3.5(2) they are special boundary points of $\mathcal{S}_1$.

(3) Again by Theorem 3.5(2) these special boundary points are interior points of $\mathcal{S}_0$. Moreover $\mathcal{C}_k$ lies in cell k, $\mathcal{C}_0 \setminus \mathcal{C}_k$ outside the cell. Thus $\mathcal{S}_1$ and $\mathcal{S}_0 \setminus \mathcal{S}_1$ connect only on the boundary of the cell, i.e., in one of the two special boundary points of $\mathcal{S}_1$.

(4) The H-index sum for $\vartheta = \pi/2$ is

$$i_H \mathcal{W}(r_k, \pi/2) + i_H \mathcal{W}(r_{k+1}, \pi/2) = i_H \mathcal{W}(r_k, \pi/2)(1+(-1)^\beta),$$

where β by Theorem 3.5 exceeds the cardinality of both N_k and N_k^*, respectively, by 1, because β also counts the kth zero of J_ν. If the cardinality of one of these sets is odd, the H-index sum is nontrivial and thus different from the H-index sum for the other interval boundary.

By Theorem 1.1 we get the existence of a bifurcating branch $\mathscr{C}$ of zeros of F^H, which satisfies a Rabinowitz type alternative and shares no interior point with $\mathscr{S}_0$. As a consequence it is contained in one of the following sets:

$$\mathscr{K}_m := \{(\vartheta, R, u) | R > 0 \text{ and } m\pi + \pi/2 < \vartheta < (m+1)\pi + \pi/2\}.$$

This holds, since the intersection $\mathscr{K}_m \cap \mathscr{C}$ is both closed and open in $\mathscr{C}$, the latter because points on the boundary of $\mathscr{K}_m$ are interior points of $\mathscr{S}_0$ and cannot be contained in $\mathscr{C}$. Thus we get $\mathscr{K}_m \cap \mathscr{C} = \mathscr{C}$.

Due to the remarks preceding this proof, $\mathscr{C} \subseteq \mathscr{K}_m$ implies that points on $\mathscr{C}$ are classical solutions of problem (12). Finally, because of the continuity of the embedding $C^2(\overline{\mathbb{B}_1(0)}) \hookrightarrow W^{2,p}(\mathbb{B}_1(0))$, the C^2-norm is unbounded, if the $W^{2,p}$-norm is. This completes the proof. □

LEMMA 3.7. *If $\mathcal{W}(R, \vartheta)$ is not an interior point of $\mathscr{S}_0$, then we have $R \geqq R_1(\vartheta)$. If equality and furthermore if $w_\vartheta(R, \vartheta) \neq 0$ holds, then the kernel of $DF^{O(n-1)}(\mathcal{W}(R, \vartheta))$ is spanned by*

$$\mathcal{W}_R(R, \vartheta) \quad \text{and} \quad \frac{\partial}{\partial x_n} \mathcal{W}(R, \vartheta)(\|x\|).$$

Proof. Elements in $\mathscr{S}_0$ with $\vartheta \equiv \pi/2$ are interior points by Theorem 3.5. So we can restrict to the case $\vartheta \not\equiv \pi/2$. Already knowing the radially symmetric elements in the kernel of DF by Lemma 3.6, we subtract from any other element $(\rho, S, u) \in \operatorname{Kern} DF(\mathcal{W}(R, \vartheta))$ in the kernel its "symmetric part" $(\rho, S, v) \in \operatorname{Kern} DF^{O(n)}(\vartheta, R, w)$, where

$$v := \int_{O(n)} \gamma \cdot u \, d\gamma.$$

The difference $u - v$ solves the differential equation

$$\Delta(u-v) + R^2 g_u(\cos\vartheta, w(R\cdot))(u-v) = 0.$$

Passing to polar coordinates $(r, \xi) \in \mathbb{R} \times S^{n-1}$ we expand $(u-v)$ in a series

$$(u-v)(r, \xi) = \sum_{\kappa=1}^{\infty} a_\kappa(r) \Phi_\kappa(\xi).$$

Here Φ_κ denote eigenfunctions of the spherical Laplacian for the eigenvalue $-\kappa(\kappa+n-2)$ and the functions $a_\kappa(r) := \langle (u-v)(r, \xi), \Phi_\kappa(\xi) \rangle_{L_2(S^{n-1})}$, $0 < r < 1$ are solutions of a singular Sturm-Liouville problem, which is proved in [15, pp. 418-420] to have (up to constant multiple) a unique solution and satisfies a Sturm-Liouville-type comparison principle. As a consequence we have $a_1(s) = \text{const.} \cdot Rw_r(R \cdot s, \vartheta)$, since $Rw_r(R \cdot s, \vartheta)$ satisfies the same equation as a_1 does. Moreover, if $a_\kappa(1) = 0$ for some $\kappa > 1$, then a_1 must have a zero in the interior of the interval $[0, 1]$, which implies $R_1(\vartheta) < R$.

If $R_1(\vartheta) > R$, all elements in the kernel therefore are radially symmetric and the dimension of the kernel is 1 by Lemma 3.6. The implicit function theorem applies and $\mathcal{W}(R, \vartheta)$ is an interior point of $\mathscr{S}_0$.

If $R_1(\vartheta) = R$ and $w_\vartheta(R, \vartheta) \neq 0$, then by Lemma 3.6 the symmetric part of the kernel is spanned by $\mathcal{W}_R(R, \vartheta)$, whereas—using Lemma 3.4—the generator for the nonsymmetric part is given by

$$\text{const. } a_1(\|x\|)\Phi_1(\xi) = w_r(R \cdot \|x\|, \vartheta)\frac{x_n}{\|x\|} = \frac{\partial}{\partial x_n}\mathcal{W}(R, \vartheta)(\|x\|). \qquad \square$$

Proof of Theorem 3.2(1). In analogy to the proof of Theorem 3.1 this proof is an application of Corollary 1.2 in § 1. We only have to check the assumptions for Corollary 1.2.

Due to Lemma 3.6 the implicit function theorem can be applied to $F^{O(n)}$ to obtain a local parametrization $R \mapsto (\vartheta(R), R, u(R))$ of the zeros of $F^{O(n)}$ with tangent vector $\mathcal{W}_R(R_1, \vartheta_1)$. Lemma 3.7 now yields the desired splitting of the kernel

$$\ker DF^H(\mathcal{W}(R_1, \vartheta_1)) = \mathbb{R} \cdot \mathcal{W}_R(R_1, \vartheta_1) \oplus \mathbb{R} \cdot \mathcal{W}_{x_n}(R_1, \vartheta_1).$$

It remains to check the transversality condition. In order to increase the clarity of the following calculations we introduce the following notation:

$$G(R, x) := R^2 g_u(\cos \vartheta(R), w(R\|x\|, \vartheta(R))),$$

$$u(R, x) := \partial_{x_n}\mathcal{W}_0(R, \vartheta(R))(x) = R\frac{x_n}{\|x\|}w_r(R\|x\|, \vartheta(R)).$$

Using this notation, we get

$$D^2F^H(\mathcal{W}(R_1, \vartheta_1))(\mathcal{W}_R(R_1, \vartheta_1), \mathcal{W}_{x_n}(R_1, \vartheta_1)) = (0, G_R(R_1)u(R_1)).$$

Suppose there is a $(\rho, S, v) \in \mathbb{R}^2 \times X$ with

$$(0, G(R_1)u(R_1)) = DF(\mathcal{W}(R_1, \vartheta_1))(\rho, S, v),$$

then integrating over $\gamma \in O(n)$ yields

$$\begin{aligned} DF(\mathcal{W}(R_1, \vartheta_1))(\rho, S, \tilde{v}) &= DF(\mathcal{W}(R_1, \vartheta_1))\left(\rho, S, \int_{O(n)} \gamma v \, d\gamma\right) \\ &= \left(0, \int_{O(n)} \gamma G(R_1)u(R_1) \, d\gamma\right) \\ &= \left(0, G(R_1)\frac{R_1}{\|x\|}w_r(R_1\|x\|, \vartheta_1)\int_{O(n)} \gamma^{-1}x_n \, d\gamma\right) \\ &= 0, \end{aligned}$$

since $\int \gamma^{-1}x_n \, d\gamma = 0$ is an $O(n)$-invariant element of $\mathbb{R}^n$. Thus $(\rho, S, \tilde{v})$ lies in the kernel of $DF^{O(n)}$ and we have

$$\begin{aligned} (0, G_R(R_1)u(R_1)) &= DF(\mathcal{W}(R_1, \vartheta_1))((\rho, S, v) - (\rho, S, \tilde{v})) \\ &= (v(0) - \tilde{v}(0), (\Delta + G(R_1))(v - \tilde{v})). \end{aligned}$$

Using the L_2 scalar product results in

$$\begin{aligned} \int_{\mathbb{B}(0)} G_R(R_1)u^2(R_1) \, dx &= \int_{\mathbb{B}(0)} ((\Delta + G(R_1))(v - \tilde{v}))u(R_1) \, dx \\ &= \int_{\mathbb{B}(0)} (v - \tilde{v})(\Delta u(R_1) + G(R_1)u(R_1)) \, dx = 0, \end{aligned}$$

since $\mathcal{W}_{x_n}(R_1, \vartheta_1)(x) = (0, 0, u(R_1, x)) \in \ker DF^H(\mathcal{W}(R_1, \vartheta_1))$. On the other hand, taking the derivative of $\int_{\mathbb{B}(0)} (\Delta u + Gu)u\,dx$ with respect to R at R_1, we arrive at

$$\int_{\mathbb{B}(0)} G_R u^2\,dx = -\int_{\mathbb{B}(0)} \underbrace{(\Delta u + Gu)}_{=0} u_R\,dx - \int_{\mathbb{B}(0)} (\Delta u_R + Gu_R)u\,dx$$

$$= -\int_{\mathbb{B}(0)} u_R \underbrace{(\Delta u + Gu)}_{=0}\,dx - \int_{\partial\mathbb{B}(0)} \left(\underbrace{\frac{d}{dn} u_R \cdot u}_{=0} - u_R \cdot \frac{d}{dn} u \right) d\omega$$

$$= R_1^3 [w_{rr}(R_1, \vartheta_1)]^2 \int_{\partial\mathbb{B}(0)} x_n^2\,dw > 0.$$

This is a contradiction. □

Appendix. By Theorem 3.1 we get bifurcation from the subset $\mathcal{C}_k \subseteq \mathcal{C}_0$ of the component $\mathcal{C}_0$ established in Theorem 2.2, if one of the following sets has odd cardinality:

$$N_k := \{[J_\nu]_k < R < [J_\nu]_{k+1} | J_{\nu+\kappa}(R) = 0 \text{ for some } 0 \leqq \kappa \in \mathbb{Z}\},$$

$$N_k^* := \{[J_\nu]_k < R < [J_\nu]_{k+1} | J_{\nu+k}(R) = 0 \text{ for some } 0 \leqq \kappa \equiv 0(2)\}.$$

We present these cardinalities in Tables 1-5 for several values of the space dimension n, which results in order $\nu = (n-2)/2$ for the Bessel function primarily considered.

TABLE 1
$n = 2$, $\nu = 0$

Zeros $[J_\nu]_k$	Absolute error	$\lvert N_k \rvert$	$\lvert N_k^* \rvert$
2.404825	5.064483E−07	2	1
5.520079	1.357311E−06	4	2
8.653714	2.926365E−05	7	3
11.791553	3.997591E−05	10	5
14.930917	5.292858E−05	11	5
18.071055	8.392533E−05	15	8
21.211600	1.947945E−04	17	7
24.352527	2.237075E−04	17	9
27.493476	2.529771E−04	23	12
30.634589	2.836580E−04	23	10

TABLE 2
$n = 3$, $\nu = 0.5$

Zeros $[J_\nu]_k$	Absolute error	$\lvert N_k \rvert$	$\lvert N_k^* \rvert$
3.141590	6.001238E−06	2	1
6.283191	1.291279E−05	5	2
9.424749	7.858169E−05	6	3
12.566098	7.284342E−04	10	5
15.708351	9.147864E−04	11	5
18.849543	1.128309E−03	15	8
21.991100	1.449951E−03	18	8
25.132608	2.075585E−03	19	10
28.274045	3.378452E−03	21	10

TABLE 3
$n=4,\ \nu=1.0$

Zeros $[J_\nu]_k$	Absolute error	$\lvert N_k\rvert$	$\lvert N_k^*\rvert$
3.831691	3.010758E−05	2	1
7.015613	5.812487E−05	5	2
10.173414	1.977336E−04	6	3
13.323294	1.202979E−03	10	5
16.471210	1.494357E−03	?	?
19.615840	1.824869E−03	?	?
22.760020	2.297772E−03	18	8
25.903509	3.146578E−03	19	10
29.047116	3.529324E−03	21	10

TABLE 4
$n=5,\ \nu=1.5$

Zeros $[J_\nu]_k$	Absolute error	$\lvert N_k\rvert$	$\lvert N_k^*\rvert$
4.493361	9.678365E−05	2	1
7.725331	1.742361E−04	5	2
10.904028	4.432954E−04	7	4
14.066321	5.729008E−04	9	4
17.220750	7.128862E−04	13	6
20.371277	9.078495E−04	13	7
23.519371	1.286108E−03	18	8
26.665856	2.121918E−03	19	10
29.811983	2.373509E−03	23	10

TABLE 5
$n=6,\ \nu=2.0$

Zeros $[J_\nu]_k$	Absolute error	$\lvert N_k\rvert$	$\lvert N_k^*\rvert$
5.135503	2.380439E−04	2	1
8.417431	4.075584E−04	5	2
11.619694	8.814832E−04	7	4
14.796154	1.124353E−03	10	4
17.959812	1.382462E−03	13	7
21.116963	1.714914E−03	?	6
24.270010	2.278524E−03	?	8
27.420746	2.574711E−03	21	11
30.569185	2.872035E−03	21	9

The zeros were computed using a scheme of Cayley and Raleigh [17, p. 502]. The scheme produces upper and lower bounds for the product of the first κ zeros. Comparing this result for k and $k-1$ upper and lower bounds can be derived for the kth zero. Sometimes these bounds were not good enough to reliably compare the zeros of two different Bessel functions. If this was the case all table entries affected by this comparison were marked with a question mark.

REFERENCES

[1] G. CERAMI, *Symmetry for a class of semilinear elliptic problems*, Nonlinear Anal., 10 (1986), pp. 1–14.

[2] G. CICOGNA, *Bifurcation and symmetries*, Boll. Un. Mat. Ital. B (6), 1 (1982), pp. 787–796.

[3] ———, *Bifurcation from topology and symmetry arguments*, Boll. Un. Mat. Ital. A(6), 3 (1984), pp. 131–138.

[4] P. CLEMENT AND G. SWEERS, *Existence and multiplicity results for a semilinear elliptic eigenvalue problem*, Tech. Report 86-51, Delft University of Technology, Delft, the Netherlands, 1986.

[5] M. G. CRANDALL AND P. H. RABINOWITZ, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.

[6] G. EISENACK AND C. FENSKE, *Fixpunkttheorie*, B.I., Mannheim, Zürich, 1978.

[7] T. KATO, *Perturbation Theory for Linear Operators*, Second edition, Springer-Verlag, Berlin, New York, 1980.

[8] P. L. LIONS, *On the existence of positive solutions of semilinear equations in* $\mathbb{R}^n$, SIAM Rev., 24 (1982), pp. 441–467.

[9] C. POSPIECH, *Global bifurcation with symmetry breaking*, Ph.D. thesis, Ruprecht-Karls-Universität, Heidelberg, 1988.

[10] P. H. RABINOWITZ, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.

[11] R. SCHAAF, *Existenz, Gestalt und Stabilität von stationären Lösungen einer Klasse von nicht-linearen parabolischen Differentialgleichungen*, Ph.D. thesis, Ruprecht-Karls-Universität, Heidelberg, 1980.

[12] C. G. SIMADER, *On Dirichlet's Boundary Value Problem*, Springer-Verlag, Berlin, New York, 1972.

[13] J. A. SMOLLER AND A. G. WASSERMAN, *Existence, uniqueness and nondegeneracy of positive solutions of semilinear elliptic equations*, Comm. Math. Phys., 95 (1984), pp. 129–159.

[14] ———, *Symmetry-breaking for positive solutions of semilinear elliptic equations*, Arch. Rational Mech. Anal., 95 (1986), pp. 217–225.

[15] ———, *Symmetry-breaking for solutions of semilinear elliptic equations with general boundary conditions*, Comm. Math. Phys., 105 (1986), pp. 415–441.

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

[17] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Second edition, Cambridge University Press, London, 1966.

[18] G. T. WHYBURN, *Topological Analysis*, Second edition, Princeton University Press, Princeton, NJ, 1964.

[19] E. ZEIDLER, *Nonlinear Functional Analysis and Its Applications*, Springer-Verlag, Berlin, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1453-1486, November 1990
© 1990 Society for Industrial and Applied Mathematics
006

BIFURCATIONS OF RELATIVE EQUILIBRIA*

MARTIN KRUPA†

Abstract. This paper discusses the dynamics and bifurcation theory of equivariant dynamical systems near *relative equilibria*, that is, group orbits invariant under the flow of an equivariant vector field. The theory developed here applies, in particular, to secondary steady-state bifurcations from invariant equilibria. Let Γ be a compact group of symmetries of R^n and let x_0 be in R^n. Suppose that f is a smooth Γ-equivariant vector field and Σ the isotropy group of x_0. It is shown that there exists a Σ-equivariant vector field f_N, defined on the space normal to X at x_0, and that the local asymptotic dynamics of f are closely related to the local asymptotic dynamics of f_N. Next those bifurcations of X are studied which occur when an eigenvalue of $(df_N)_x$ crosses the imaginary axis. Properties of the vector field f_N imply that branches of equilibria and periodic orbits of f_N correspond to trajectories of f which are dense in tori. Field [*Equivariant dynamical systems*, Trans. Amer. Math. Soc., 259 (1980), pp. 185-205] found bounds on the dimensions of these tori. Some of his results are extended. This theory is applied to the following specific problems:

(1) Bifurcations of systems with $O(2)$ symmetry.
(2) Bifurcations of steady-state solutions of the Kuramoto-Sivashinsky equation.
(3) Secondary bifurcations in the planar Bénard problem.

Key words. bifurcation, symmetry, relative equilibria

AMS(MOS) subject classifications. 58F14, 58F27, 34C35

Introduction. In this work we discuss the dynamics and bifurcation theory of equivariant dynamical systems near group orbits invariant under the action of the flow. Such group orbits are called *relative equilibria.* The simplest example of a relative equilibrium is a group orbit of equilibria. A group orbit of equilibria can be characterized as a relative equilibrium on which the flow is trivial. The symmetry groups we consider are compact and have positive dimension, so, in particular, they must contain a subgroup isomorphic to $SO(2)$. For such groups the flow trajectories on the relative equilibria can be nontrivial. A well-known example of such nontrivial trajectories on relative equilibria are rotating waves, that is, solutions given by $x(t) = \theta(t)x_0$, where $\theta(t)$ parametrizes $SO(2)$.

A special case of a relative equilibrium is an invariant equilibrium, that is, an equilibrium invariant under all the symmetries of the system. Such equilibria often arise in applications, and their bifurcations have been extensively studied. Often bifurcations of invariant equilibria are characterized by symmetry breaking; that is, the invariant equilibrium bifurcates to branches of equilibria no longer invariant under the action of the symmetry group. In other words, nontrivial relative equilibria often occur as a result of bifurcations of invariant equilibria. In this context bifurcations of relative equilibria correspond to secondary bifurcations from an invariant equilibrium.

Let X be a group orbit of equilibria of an equivariant vector field f. If X has positive dimensions, then the conditions determining when f can undergo a bifurcation near X are quite different than in the case of an invariant equilibrium. In particular, no element of X can be hyperbolic, since the directions along the group orbit must be neutrally stable. More precisely, for any $x \in X$ the tangent space T_xX is contained in the kernel of the derivative $(df)_x$. It follows that X will be normally hyperbolic if

* Received by the editors July 19, 1989; accepted for publication (in revised form) December 1, 1989. This research was supported in part by National Science Foundation/Defense Advanced Research Projects Agency grant DMS-8700897 and NASA Ames grant NAG-2-432.

† Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455.

$(df)_x$ has no purely imaginary eigenvalues and the algebraic multiplicity of zero as an eigenvalue of $(df)_x$ equals the dimension of T_xX. A bifurcation of X will occur if $(df)_x$ has an eigenvalue on the imaginary axis whose (generalized) eigenvector is not contained in the tangent space of X.

An interesting phenomenon has been observed in examples of bifurcations of relative equilibria: an orbit of equilibria can lose stability by having an eigenvalue passing through zero and bifurcate to a group orbit consisting of nontrivial flow trajectories. The flow on the new relative equilibria is a slow drift given by the action of a curve of group elements. Several authors, who studied bifurcations of relative equilibria, found that the resulting dynamics could be described in terms of dynamics related to standard bifurcations modulated by a drift along the group orbit. The following articles focused on bifurcations of relative equilibria where this feature has been observed. Chossat [1986] has shown that the bifurcation of standing waves in the problem of degenerate Hopf bifurcation with $O(2)$ symmetry leads to quasi-periodic motion on a group invariant two-dimensional torus. Iooss [1986] has shown that a Hopf-Hopf mode interaction in the Taylor-Couette problem ($O(2) \times SO(2)$ symmetry) leads to a three-frequency flow. Danglemayr [1986] has found a rotating wave in the problem of steady-state mode interaction with $O(2)$ symmetry. Chossat and Golubitsky [1988] have studied a related problem of Hopf bifurcation of a group orbit of standing waves and have discovered that this bifurcation leads to a three-frequency motion, with one of the frequencies given by the drift along the orbit. In their paper Chossat and Golubitsky have formulated the following theorem: the flow near a relative equilibrium can be decomposed into the flow in the direction along the orbit and the flow in the direction normal to the orbit. The precise statement and the proof of this theorem is the starting point of this work.

Section 1 of the paper contains some background information on Lie group theory. The remaining part of the paper is divided into two parts. The first part, §§ 2–5, is devoted to the theoretical aspects of the problem. The second part, §§ 6–8, focuses on specific group actions and specific dynamical systems and is designed to show the application of the ideas developed in the first part. The reader more interested in the second part of the paper will only need to know the definitions and the statements of theorems contained in the first part. The following is a brief description of the topics discussed in each of the sections.

In § 2 we give a precise description of how the previously mentioned decomposition of the vector field can be accomplished. We show that near relative equilibria the vector field can be written as a sum of equivariant components: one tangent to the group orbits and the other normal to the original orbit X (Theorem 2.1). As a consequence of this decomposition each bounded solution near a relative equilibrium is contained in the group orbit of a solution of the normal vector field f_N (Theorem 2.2). In the remainder of § 2 we show that the asymptotic dynamics of f can be determined by the asymptotic dynamics of the normal vector field modulo drifts along the orbit. Some results of § 2, including an alternative proof of Theorem 2.1, can be found in Vanderbauwhede, Krupa, and Golubitsky [1989].

The results of § 2 imply that bifurcations of f can be analyzed in two steps. The first step is to describe bifurcations of the normal vector field f_N and the second step is to find the corresponding drifts along group orbits. Let x be in X. In § 3 we argue that generic bifurcations of f_N can be described as bifurcations of a generic Σ_x-equivariant vector field, where Σ_x is the isotropy subgroup of x.

Suppose that f describes a family of vector fields, rather than a single vector field. In § 4 we study bifurcations of relative equilibria occurring when an eigenvalue of

$(df_N)_x$ passes through zero. We analyze the case when a relative equilibrium X bifurcates to another relative equilibrium Y. Let $y \in Y$, let Σ be the isotropy subgroup of y and $N(\Sigma)$ the normalizer of Σ. Field [1980] proves a theorem stating that the flow on Y is given by a linear flow on a torus whose dimension is bounded by, and generically equal to, rank $(N(\Sigma)/\Sigma)$ (rank $(N(\Sigma)/\Sigma)$ equals the dimension of a maximal torus in $N(\Sigma)/\Sigma$). The main theorem of § 4 (Theorem 4.1′) states that there exists a generic set of perturbations of f whose elements have the following property: for all except countably many values of the parameter the dimension of the flow on Y is maximal.

In § 5 we study Hopf bifurcations of relative equilibria, that is, bifurcations occurring when an eigenvalue of $(df_N)_x$ passes through a nonzero point on the imaginary axis. We apply the standard Hopf bifurcation theorems to find periodic solutions of the normal component f_N. Let Y be a periodic orbit of f_N and let Σ be the isotropy subgroup of the elements of Y. Field [1980] shows that the corresponding trajectories of f are dense in tori whose dimension is bounded by rank $(N(\Sigma)/\Sigma)+1$. Let $\tilde{\Sigma}$ be the group consisting of all the symmetries that leave Y invariant. In Theorem 5.1 we derive a new bound, given by rank $(N(\tilde{\Sigma})/\tilde{\Sigma})+1$ and show that this bound is attained for a generic vector field. Next, in Theorem 5.2, we consider a family of vector fields f, such that f_N has a Hopf bifurcation and show that there exists a generic set of perturbations of f whose elements are such that for all except countably many values of the parameter the dimension of the flow on the manifolds ΓY is maximal.

In § 6 we present a classification of generic secondary steady-state and Hopf bifurcations with symmetry group $O(2)$. In this context bifurcations of the normal vector field correspond to steady-state and Hopf bifurcations with D_k symmetry. Using the results of §§ 3 and 4 we determine for which bifurcations of the vector field f_N the bifurcating solutions of the full vector field f generically have nontrivial drift along group orbits.

In § 7 we analyze bifurcations of the zero solution of the Kuramoto–Shivashinsky equation, which has $O(2)$ symmetry. We summarize the results of a computer-assisted study done by Kevrekedis, Nicolaenco, and Scovel [1988] and compare their numerical results with the predictions of $O(2)$ bifurcations, as described in § 6.

In § 8 we classify the possible generic steady-state bifurcations in the planar Bénard problem. The generic primary bifurcations in the Bénard problem are to two types of equilibria: hexagons (with symmetry D_6) and rolls (with symmetry $O(2)\oplus Z_2$). We consider secondary steady-state bifurcations of hexagons and rolls and show that the resulting trajectories are either equilibria or rotating waves.

1. Preliminaries. Let Γ be a compact Lie group. We consider a smooth linear action of Γ on R^n. With no loss of generality we can assume that this action is orthogonal and hence identify Γ with a subgroup of $O(n)$ (see Bredon [1972, I, 3.5]). Let X be a compact and Γ-invariant submanifold of R^n. For $x \in X$ let N_x be the set of vectors normal to X. Note that N_x is a vector subspace of R^n, since it passes through zero. Let $N(X)$ be the bundle with base space X and fibers N_x; $N(X)$ is called the *normal bundle* of X. The bundle $N(X)$ is smooth (see Guillemin and Pollack [1974, p. 71]). The action of Γ on $N(X)$ is defined by the formula $\gamma(x, v) = (\gamma x, \gamma v)$. To see that this action is well defined observe that the orthogonality of the action of Γ implies that $\gamma N_x = N_{\gamma x}$. Let $\beta : N(X) \to R^n$ be defined as $\beta((x, u)) = x + u$. It is easy to see that the map β is Γ-equivariant and a local diffeomorphism. It follows that an invariant neighborhood of X in R^n can be identified, via the map β, with a neighborhood of the zero section in $N(X)$. For $x \in X$ let $\hat{N}_x = \{(x, v) : v \in N_x\}$. Note that $\hat{N}_x \subset N(X)$

and the image of $\hat{N}_x$ under β is in N_x. Given a Γ-equivariant vector field $f: R^n \to R^n$ let $\beta^* f$ be defined by $\beta^* f(y) = [d\beta_{\beta(y)}]^{-1} f(\beta(y))$, $y \in N(X)$. The map $\beta^* f$ is called the *pullback* of f to $N(X)$. Let $x \in R^n$ and $X = \Gamma x$. It follows that studying local dynamics of f near x is equivalent to studying the dynamics of its pullback to $N(X)$.

For $x \in R^n$ (or $N(X)$) let

$$\Sigma_x = \{\sigma \in \Gamma: \sigma x = x\}$$

be the *isotropy subgroup* of x.

The homogeneous space Γ/Σ_x is not, in general, a group, but it has the structure of a smooth manifold and the quotient map $\pi: \gamma \mapsto \gamma\Sigma_x$ is a surjection (see Bredon [1972, p. 302]). The map $\gamma\Sigma_x \mapsto \gamma x$ is a diffeomorphism between Γ/Σ_x and the orbit Γx of x (see Bredon [1972; VI, 1.2]). Fix $x \in R^n$ and let $\Sigma = \Sigma_x$. Suppose that there exists a neighborhood U of $e\Sigma$ in Γ/Σ and a map $\sigma: U \to \Gamma$ such that $\pi\sigma(u) = u$ for all $u \in U$. The map σ is called a *local cross section* of π (for a more general definition see Bredon [1972, p. 39]). We now construct a local cross section of π. Let Δ be a submanifold of Γ transverse to Σ at e with $e \in \Sigma$ and $\dim \Sigma + \dim \Delta = \dim \Gamma$. Note that a neighborhood $\hat{U}$ of e in Δ is diffeomorphic to a neighborhood of $e\Sigma$ in Γ/Σ and this diffeomorphism is given by $\pi|\hat{U}$. Let $\sigma = (\pi|\hat{U})^{-1}$. Clearly, the map σ is a local cross section of π.

Let σ be a cross section of π defined on a neighborhood U of $e\Sigma$ in Γ/Σ. A simple argument shows that the map $\phi: \sigma(U) \times N_x \to R^n$ defined as $\phi(u, y) = \sigma(u)y$ is a local diffeomorphism. Let N_x^ε be a disc of radius ε around x and let ε be chosen so that $\phi: \sigma(U) \times N_x^\varepsilon$ is a diffeomorphism. Let $V^\varepsilon = \Gamma N_x^\varepsilon$. Orthogonality of the action implies that N_x and N_x^ε are Σ_x-invariant. Let $X = \Gamma x$. We chose ε so that the set V^ε is equivariantly diffeomorphic to a neighborhood of the zero section in $N(X)$. Observe that if $y \in \hat{N}_x$ then it is clear that $\Sigma_y \subset \Sigma_x$. Hence if $y \in N_x^\varepsilon$ then $\Sigma_y \subset \Sigma_x$. We have the following proposition.

PROPOSITION 1.1. *Every smooth and Σ_x-equivariant vector field $g: N_x^\varepsilon \to R^n$ has a unique smooth and Γ-equivariant extension $f: V^\varepsilon \to R^n$.*

Proof. We define f by requiring that $f(\gamma y) = \gamma g(y)$ for $\gamma \in \Gamma$, $y \in N_x^\varepsilon$. To see that f is well defined let $\gamma_1 y = \gamma_2 y$, $y \in N_x^\varepsilon$, $\gamma_1, \gamma_2 \in \Gamma$. Then $\gamma_1^{-1}\gamma_2 y = y$, so $\gamma_1^{-1}\gamma_2 \in \Sigma_y \subset \Sigma_x$ and $g(\gamma_1^{-1}\gamma_2 y) = g(y)$, implying $\gamma_1 g(y) = \gamma_2 g(y)$. Hence f is well defined.

Let U, σ, and ϕ be as defined prior to the statement of Proposition 1.1. Let $\hat{U} = \phi(U)$. Then $f|\hat{U} = \phi \circ id \times h \circ \phi^{-1}$. It follows that f is smooth on $\hat{U}$. Smoothness of f on V^ε follows from equivariance and smoothness of the action.

2. Dynamics near relative equilibria. Let Γ be a Lie subgroup of $O(n)$ acting orthogonally on R^n and let $f: R^n \to R^n$ be a C^∞ smooth Γ-equivariant vector field. Fix x_0 in R^n and let X denote the group orbit of x_0. We say that the set X is a *relative equilibrium* of f if X is invariant under the flow of f. The subject of this work is to study bifurcations of relative equilibria. In this section we develop a systematic way of analyzing dynamics near a relative equilibrium X. We first describe our results, deferring the proofs to the end of the section.

We begin by defining the concepts of a tangent vector field and a normal vector field. Let $g: R^n \to R^n$. We say that g is a *tangent vector field* if $g(u)$ is tangent to the group orbit of x for all x in R^n. For x in X let N_x be the space of vectors normal to X at x. We say that g is a *normal vector field* if for every x in X the space N_x is invariant under the flow of g. Note that a normal vector field does not have to be normal to group orbits other than X.

This section contains two main theorems. The first theorem states that near the group orbit X the vector field f can be written as a sum of a smooth Γ-equivariant

normal vector field f_N and a smooth Γ-equivariant tangent vector field f_T. We refer to this theorem as the decomposition theorem. The second theorem states that near X the dynamics of f can be described as dynamics of f_N modulated by drifts along group orbits.

The decomposition theorem is a consequence of a technical lemma. The lemma states that near X there exists a smooth, Γ-invariant bundle K whose fibers are tangent to group orbits and whose restriction to X is the tangent bundle of X. Before stating the lemma we discuss the concept of the normal bundle of X. We observe that a Γ-invariant neighborhood of X can be identified with a neighborhood of the zero section in the normal bundle of X. We conclude that the dynamics of f can be understood in terms of the dynamics of the pullback of f to the normal bundle. In the proof of the decomposition theorem we identify f with its pullback to the normal bundle.

Following the proof of the decomposition theorem we discuss the most important implications of the two main theorems. We remark that Γ-equivariance of f_N implies that the dynamics of f_N is completely determined by its dynamics on the invariant set N_{x_0}. We also describe a way of finding a global center manifold near a relative equilibrium X. When X is an orbit of equilibria we show that the global center manifold is the union of local center manifolds constructed for each normal space N_x.

Next we discuss a method of explicit computation of f_N. We present the general form of "coordinates along group orbits" and "in the direction normal to the orbit." Such coordinates have been used to study specific examples of bifurcations of relative equilibria.

We begin by assuming that $X = \Gamma x_0$ is the group orbit of x_0, but not necessarily a relative equilibrium of f. The following theorem is the first of the two main results of this section.

THEOREM 2.1. *There exists a Γ-invariant neighborhood U of X in R^n, a smooth and Γ-equivariant normal vector field f_N, and a smooth and Γ-equivariant tangent vector field f_T such that*

$$f(u) = f_T(u) + f_N(u)$$

for all u in U.

Let g denote the restriction of f_N to the space N_{x_0}. Let U be the neighborhood defined in Theorem 2.1 and suppose that $u(t)$ is a trajectory of f contained in U for all $t \geqq 0$. We now state the second of the two main theorems of this section.

THEOREM 2.2. *There exists a smooth curve of group elements $\gamma(t)$ and a trajectory $y(t)$ of the vector field g such that $\gamma(t)y(t) = u(t)$ for all $t \geqq 0$.*

Let $\Pi: N(X) \to X$ be the bundle projection, that is, $\Pi((x, v)) = x$. The following lemma is the main technical result necessary to prove Theorem 2.1.

LEMMA 2.3. *There exists a smooth Γ-invariant subbundle K of $TN(X)$ such that for all $y \in N(X)$*

(i) $K_y \subset T_y \Gamma y$

(ii) $K_y \oplus N_{\Pi(y)} = R^n$.

Note that we cannot define K_y as $T_y \Gamma y$ since the dimension of group orbits may increase near x_0 (the fact that it cannot decrease is a consequence of the inclusion $\Sigma_y \subset \Sigma_{\Pi(y)}$ for $y \in \hat{N}_{\Pi(y)}$). In fact, proving Lemma 2.3 is the main technical difficulty of this section. We defer the proof to the end of the section. The proof of Theorem 2.2 is also deferred, since it relies on the proof of Lemma 2.3. The proof of Theorem 2.1 is a simple consequence of Lemma 2.3. In the proof we assume that f is a vector field on $N(X)$; that is, $f: N(X) \to TN(X)$.

Proof of Theorem 2.1. Suppose $y \in N(X)$. Let $P: TN(X) \to TN(X)$ be defined by $P(y, v) = P_y v$, where P_y is a projection with $\ker P_y = K_y$ and $\operatorname{Im} P_y = N_{\Pi(y)}$. The map P is smooth since the spaces $\ker P_y$ and $\operatorname{Im} P_y$ vary smoothly with y. Equivariance of P follows from invariance of K and equivariance of Π. Let $f_N(y) = P(y, f(y))$, $f_T(y) = f(y) - f_N(y)$. As required f_N is a normal vector field, f_T is a tangent vector field, and they are both smooth and Γ-equivariant.

We now discuss some implications of the two main theorems. Recall that orthogonality of the action implies that N_{x_0} is Σ_{x_0}-invariant. Let g be the vector field defined following the statement of Theorem 2.1; that is, g is the restriction of f_N to the space N_{x_0}. Since f_N is Γ-equivariant it follows that g is Σ_{x_0}-equivariant. Let $k = \operatorname{codim} X$. Note that $\dim N_{x_0} = k$. It follows that in order to understand the dynamics of f near X we need to carry out two steps:

(a) Analyze the dynamics of the k-dimensional Σ_{x_0}-equivariant vector field g.

(b) Find the drift along group orbits $\gamma(t)$.

Suppose that X is a relative equilibrium; in this case every x in X is an equilibrium of f_N. In particular, the point x_0 is an equilibrium of g. Let m be a positive integer. The equivariant center manifold theorem (cf. Ruelle [1973, Thm. 1.2]) implies that near x_0 the vector field g has a C^m smooth Σ_{x_0}-invariant center manifold. Let M_{x_0} denote such a center manifold. Let $M = \Gamma M_{x_0}$. The smoothness of the action and Σ_{x_0}-invariance of M_{x_0} together imply that M is C^m smooth. Theorem 2.2 also implies that all trajectories of f contained in a sufficiently small neighborhood of X approach M as time goes to infinity. We say that M is the *center manifold* of the relative equilibrium X for the vector field f.

When X consists of equilibria it is natural to ask whether the global center manifold M is a local center manifold for every element of X. We answer this question in the affirmative by verifying that for all x in X the tangent space to M at x equals the center subspace of $(df)_x$. Let E^c be the restriction of the tangent bundle of M to the relative equilibrium X. The bundle E^c is called the *center bundle* of X. We have the following proposition.

PROPOSITION 2.4. *Let x be in X. The fiber of the center bundle E^c at x is the center subspace of $(df)_x$.*

Proof. We first prove that $T_x X$ is contained in $\ker (df)_x$. Any vector $u \in T_x X$ can be written as $(d/ds)\gamma(s)x|_{s=0}$. We use the chain rule and the fact that $f(\gamma(s)x) = 0$ to obtain

$$(df)_x u = (df)_x \left(\frac{d}{ds} \gamma(s)x|_{s=0} \right) = \frac{d}{ds} f(\gamma(s)x)|_{s=0} = 0.$$

Hence zero is an eigenvalue of $(df)_x$ with multiplicity greater than or equal $\dim X$. Let $v \in N_x$. Theorem 2.1 implies that $(df)_x v = (df_N)_x v + (df_T)_x v$. We show that $(df_N)_x v \in N_x$ and $(df_T)_x v \in T_x X$. This implies that $(df)_x$ can be written in the form:

$$(df)_x = \begin{pmatrix} 0 & (df_T)_x \\ 0 & (df_N)_x \end{pmatrix}. \tag{2.1}$$

The proposition follows from equation (2.1), since (2.1) implies that all nonzero eigenvalues of $(df)_x$ are also eigenvalues of $(df_N)_x$.

We now prove that (2.1) is valid. The vector field f_N is Γ-equivariant and $f_N(x) = 0$. The argument presented at the beginning of this proof implies that $T_x X \subset \ker (df_N)_x$. Since $f = f_N + f_T$ it follows that $T_x X \subset \ker (df_T)_x$. Recall from the proof of Theorem 2.1 that $f_N(y) = P_y f(y)$, where P_y is a projection with $\ker (P_y) = T_y \Gamma y$ and

Im $(P_y) = N_{\Pi(y)}$. Hence by the chain rule

$$(df_N)_x = (d_y P_y f(x))_x + P_x (df)_x. \tag{2.2}$$

But $f(x) = 0$, so

$$(df_N)_x = P_x (df)_x. \tag{2.3}$$

It follows that $(df_T)_x(N_x) \subset T_x X$ and $(df_N)_x(N_x) \subset N_x$. Equation (2.1) now follows.

In many applications a bifurcation of a group orbit of equilibria X occurs when $(df)_{x_0}$ maps a vector $v \in N_{x_0}$ to $T_{x_0}X$. The vector v then becomes a generalized nullvector of $(df)_{x_0}$. We have the following proposition.

PROPOSITION 2.5. *The vector v is a null vector of $(df_N)_{x_0}$.*

Proof. Proposition 2.5 follows from identity (2.3).

In applications we need to explicitly compute the vector field f_N. This can be done by changing variables to coordinates in the normal space N_{x_0} and a complementary set of coordinates "along group orbits." Such coordinates have been used by Chossat [1986], Iooss [1986], Danglemayr [1986], and others to study bifurcations of relative equilibria. In the form presented here they were suggested by Chossat and can be found in Moutrane [1988]. In our presentation we assume that f is a vector field on the normal bundle $N(X)$; that is, $f: N(X) \to TN(X)$. For a Lie group Δ, let $\mathscr{L}(\Delta)$ denote the Lie algebra of Δ. Let $\exp: \mathscr{L}(\Gamma) \to \Gamma$ be the exponential mapping. Let $V \subset \mathscr{L}(\Gamma)$ be the orthogonal complement of $\mathscr{L}(\Sigma_{x_0})$ in $\mathscr{L}(\Gamma)$ (the space V will be defined more precisely in the proof of Lemma 2.3). Let $\theta: V \times \hat{N}_{x_0} \to N(X)$ be given by $\theta(\xi, y) = (\exp \xi)(y)$. The linear map $(d\theta)_{(0,x_0)}$ is an isomorphism, and hence θ is a local diffeomorphism. Let $h = \theta^* f$. Note that for every $y \in N(X)$ the fiber of the tangent bundle $T_y N(X)$ can be written as $T_y N(X) = V \oplus N_{\Pi(y)}$. The vector field h is defined on $V \times \hat{N}_{x_0}$ and has the following property.

PROPOSITION 2.6. *If h is written in the form $h = (h_1, h_2)$, with $h_1 \in V$ and $h_2 \in N_{x_0}$, then $h_2(0, y) = f_N(y)$ for all $y \in \hat{N}_{x_0}$.*

The proof of Proposition 2.6 relies on the proof of Lemma 2.3 and therefore will be given at the end of the section.

Proof of Lemma 2.3. Let $\mathscr{L}(\Gamma)$ denote the Lie algebra of Γ and let $\exp: \mathscr{L}(\Gamma) \to \Gamma$ be the exponential mapping. We begin by recalling two concepts related to the Lie algebra $\mathscr{L}(\Gamma)$. The *action* of $\mathscr{L}(\Gamma)$ on $N(X)$ is defined by

$$\xi y = \frac{d}{dt} (\exp t\xi) y|_{t=0} \quad \text{for } \xi \in \mathscr{L}(\Gamma), \quad y \in N(X).$$

The *adjoint action* of Γ on $\mathscr{L}(\Gamma)$ is defined by

$$\mathrm{Ad}_\gamma \xi = \frac{d}{dt} \gamma(\exp t\xi)\gamma^{-1}|_{t=0} \quad \text{for } \gamma \in \Gamma, \quad \xi \in \mathscr{L}(\Gamma).$$

Note that

$$\gamma \xi y = \mathrm{Ad}_\gamma \xi \gamma y \quad \text{for } \gamma \in \Gamma, \quad \xi \in \mathscr{L}(\Gamma), \quad y \in N(X). \tag{2.4}$$

Recall that $k = \operatorname{codim} X$. We prove that finding the bundle K is equivalent to finding a bundle E over $N(X)$ whose fibers are k-dimensional subspaces of $L(\Gamma)$ having the following property:

$$E_{\gamma y} = \mathrm{Ad}_\gamma E_y \quad \text{for } \gamma \in \Gamma. \tag{2.5}$$

Suppose that the bundle E has been found. Then we define the fiber K_y of the bundle K as the set of all images of y under the action of elements of E_y; that is,

$$K_y = \{\xi y : \xi \in E_y\}.$$

If $v \in K_y$ then (2.4) implies that $\gamma v \in K_{\gamma y}$. Hence K is Γ-invariant. We now prove that K is a smooth bundle. For $y \in N(X)$ let $\Phi_y : \Gamma \to N(X)$ be defined by $\Phi_y(\gamma) = \gamma y$, $\gamma \in \Gamma$. For $\xi \in \mathscr{L}(\Gamma)$ we have

$$\xi y = d\Phi_y(e)\xi.$$

It follows that $K_y = d\Phi_y(e)E_y$. Smoothness of K now follows from smoothness of E and from smooth dependence of Φ_y on y. The definition of the action of $\mathscr{L}(\Gamma)$ implies that the fibers K_y are tangent to group orbits.

It remains to prove the existence of the bundle E. Suppose $\langle\langle\ ,\ \rangle\rangle$ is an inner product on $\mathscr{L}(\Gamma)$ invariant with respect to the adjoint action. Such an inner product always exists for a finite-dimensional action of a compact Lie group (see, for example, Golubitsky, Stewart, and Schaeffer [1988, Prop. XI, 1.3]). Let V be the orthogonal complement of $\mathscr{L}(\Sigma_{x_0})$ taken with respect to $\langle\langle\ ,\ \rangle\rangle$; that is, $V = \mathscr{L}(\Sigma_{x_0})^{\perp}$. Let E be the bundle over $N(X)$ with $E_y = \mathrm{Ad}_\gamma V$, $y \in \gamma \hat{N}_{x_0}$. To see that E is well defined suppose that $\gamma_1 y = \gamma_2 y$, $y \in \hat{N}_{x_0}$. From the properties of the normal bundle it follows that $\gamma_1 \gamma_2^{-1} \in \Sigma_{x_0}$. It follows that $\mathrm{Ad}_{\gamma_1\gamma_2^{-1}} V = V$ or $\mathrm{Ad}_{\gamma_1} V = \mathrm{Ad}_{\gamma_2} V$, implying that $E_{\gamma_1 y} = E_{\gamma_2 y}$. Also E_y is defined for all $y \in N(X)$ since $N(X) = \bigcup_{\gamma\in\Gamma} \gamma \hat{N}_{x_0}$. Equation (2.4) is automatically satisfied for the fibers of E.

For smoothness of E let U be a neighborhood of $e\Sigma_{x_0}$ in Γ/Σ_{x_0} and let $\sigma : U \to \Gamma$ be a local cross section of π. Let $\phi : U \times N_{x_0} \to N(X)$ be given as $\phi(u, y) = \sigma(u)(x_0, y)$. The map ϕ is a local diffeomorphism near $(e\Sigma_{x_0}, 0)$. It follows that the map $\Psi : U \times N_{x_0} \times V \to E$ given by $\Psi(u, y, \xi) = (\sigma(u)y, \mathrm{Ad}_{\sigma(u)}\, \xi)$ is a local bundle diffeomorphism. This shows smoothness of E near $(x_0, 0)$. To show smoothness near $(\gamma x_0, 0)$ we use the map $\gamma\Psi$ and the relation (2.4).

Proof of Theorem 2.2. Suppose $u(0) = u_0$. Let γ_0 be the element of Γ such that $u_0 \in \gamma_0 N_{x_0}$. Let $y_0 = \gamma_0^{-1} u_0$ and let $y(t)$ be the integral curve of f_N with $y(0) = y_0$. Let $\cdot$ denote differentiation with respect to t. To prove the theorem we need to find a curve $\gamma(t)$ with $\gamma(0) = \gamma_0$ and such that

$$(\gamma(t)y(t))^{\cdot} = f(\gamma(t)y(t)). \tag{2.6}$$

The idea of the proof is to reduce (2.6) to an initial value problem on Γ. We observe that the left-hand side of (2.6) can be written as

$$\frac{d}{ds}\gamma(t+s)y(t)|_{s=0} + \gamma(t)\dot{y}(t).$$

By assumption $\dot{y}(t) = f_N(y(t))$. It follows that (2.6) can be rewritten as

$$\frac{d}{ds}\gamma(t+s)y(t)|_{s=0} = \gamma(t)f_T(y(t)). \tag{2.7}$$

Let V be the subspace of $\mathscr{L}(\Gamma)$ defined in the proof of Lemma 2.3. It follows from the proof of Lemma 2.3 and from the construction of the vector field f_T that there exists a curve $\xi(t)$ of elements of V such that $f_T(y(t)) = \xi(t)y(t)$. Equation (2.7) can now be rewritten as

$$\frac{d}{ds}\gamma(t+s)y(t)|_{s=0} = \gamma(t)\xi(t)y(t). \tag{2.8}$$

Consider the initial value problem:

$$\dot{\gamma} = \gamma(t)\xi(t), \qquad \gamma(0) = \gamma_0. \tag{2.9}$$

By standard theory of ordinary differential equations, (2.9) has a unique solution $\gamma(t)$. It is clear that if $\gamma(t)$ is a solution of (2.9) then $\gamma(t)y(t)$ satisfies (2.8). The theorem now follows.

Proof of Proposition 2.6. Let $y \in \hat{N}_{x_0}$. The proof of Lemma 2.3 and the construction of f_T imply that there exists a unique $\xi_0 \in V$ such that $f_T(y) = d/dt\ (\exp t\xi_0)y|_{t=0}$. Note that $d/dt\ (\exp t\xi_0)y|_{t=0} = (d\theta)_{(0,y)}(\xi_0, 0)$ implying

$$(d\theta)_{(0,y)}(\xi_0, 0) = f_T(y). \tag{2.10a}$$

Also $\theta|\hat{N}_{x_0} = id$. Hence

$$(d\theta)_{(0,y)}(0, f_N(y)) = f_N(y). \tag{2.10b}$$

Combining (2.10a) and (2.10b) we obtain $(d\theta)_{(0,y)}(\xi_0, f_N(y)) = f_N(y) + f_T(y) = f(y)$. Also, since $y \in \hat{N}_{x_0}$ we have $\theta^* f(y) = (d\theta)^{-1}_{(0,y)} f(y)$. Hence $h_2(0, y) = f_N(y)$.

3. Bifurcations of the normal vector field. Let $F: R^n \times R \to R^n$ be a family of vector fields and assume that $F(x_0) \in T_{x_0}X$; that is, X is a relative equilibrium. The results of § 2 imply that the dynamics of F can be described as follows: the trajectory of F with initial condition y_0 is contained in the group orbit of the trajectory of F_N with the same initial condition. We will utilize this property of the dynamics and divide the bifurcation analysis into two steps. The first step will be to analyze bifurcations of the normal vector field. Then, given a bifurcating solution of the normal vector field, say $y(t)$, we will study the dynamics of F on the set $Y = \{\gamma y(t): \gamma \in \Gamma, t \in R\}$. Note that Y is Γ-invariant and, by Theorem 2.2, it is invariant under the flow of F. This program will be carried out for two kinds of trajectories of F_N—equilibria and periodic orbits.

In this section we discuss the first part of the bifurcation analysis, that is, bifurcations of the normal vector field. Suppose that $\dim N_{x_0} = k$. We prove that generic bifurcations of F_N can be described in terms of generic bifurcations of Σ_{x_0}-equivariant vector fields on R^k. More specifically, we show that a property generic in the class of smooth, Σ_{x_0}-equivariant vector fields on R^k is also generic in the class of normal vector fields on N_{x_0}. Let $G(\cdot, \lambda)$ be the restriction of the vector field $F_N(\cdot, \lambda)$ to N_{x_0}. Let $g = G(\cdot, 0)$. Suppose that $(dg)_{x_0}$ has an eigenvalue on the imaginary axis and let E be the center subspace of $(dg)_{x_0}$. Suppose that G has a steady-state bifurcation; that is, $(dg)_{x_0}$ has a zero eigenvalue. Then we have the following proposition.

PROPOSITION 3.1. *Generically the space E equals the nullspace of $(dg)_{x_0}$ and the action of Σ_{x_0} on E is irreducible.*

Proposition 3.1 follows from Proposition 1.1 and standard results in equivariant bifurcation theory (see, for example, Golubitsky, Stewart, and Schaeffer [1988, Prop. XII, 3.4]).

Suppose that W is a subspace of R^k. We say that the action of Σ_{x_0} on W is Γ-*simple* if it is irreducible but not absolutely irreducible or if there exists a space V such that $W = V \oplus V$ and the action of Σ_{x_0} on V is absolutely irreducible.

Suppose now that G has a Hopf bifurcation; that is, $(dg)_{x_0}$ has a purely imaginary eigenvalue $i\omega$. The following proposition gives a characterization of the space E.

PROPOSITION 3.2. *Generically the space E is the generalized eigenspace of $i\omega$ for $(dg)_{x_0}$ and the action of Σ_{x_0} on E is Γ-simple.*

Proposition 3.2 follows from Proposition 1.1 and standard results in bifurcation theory (see, for example, Golubitsky, Stewart, and Schaeffer [1988, Prop. XVI, 1.4]).

A center manifold reduction coupled with a change of coordinates allows us to reduce the original bifurcation problem for G to a bifurcation problem posed on $E \times R$. We divide the analysis into two cases:

(i) The action of Σ_{x_0} on E is trivial.

(ii) The action of Σ_{x_0} on E is nontrivial.

Case (i) is much simpler and can be analyzed simultaneously for all groups Γ. In particular no symmetry breaking takes place. The following proposition summarizes the bifurcation analysis for this case.

PROPOSITION 3.3. *Suppose that the action of Σ_{x_0} on E is trivial. If $(dg)_{x_0}$ has a zero eigenvalue, then generically G has a limit point bifurcation. If $(dg)_{x_0}$ has a purely imaginary eigenvalue $i\omega$, then generically $i\omega$ is a simple eigenvalue of $(dg)_{x_0}$ and G has a Hopf bifurcation to a unique periodic solution.*

The proof of Proposition 3.3 follows from standard results in bifurcation theory. In the remainder of this work, unless otherwise stated, we will assume that the action of Σ_{x_0} on E is nontrivial.

Let $y = y(\lambda)$ be a branch of equilibria of G and $Y = Y(\lambda)$ a branch of periodic orbits of G. Let Σ be the isotropy subgroup of y and Σ_Y the group of symmetries mapping Y into itself. In §§ 4 and 5 we show that the trajectories of F on Γy are dense in tori whose dimension is bounded by rank $(N(\Sigma)/\Sigma)$ (the dimension of a maximal torus in $N(\Sigma)/\Sigma$) and the trajectories on ΓY are dense in tori whose dimension is bounded by rank $(N(\Sigma_Y)/\Sigma_Y)+1$. Generically these tori are of maximal dimension. In the context of Proposition 3.3, this maximal dimension equals rank $(N(\Sigma_{x_0})/\Sigma_{x_0})$ for trajectories on Γy and rank $(N(\Sigma_{x_0})/\Sigma_{x_0})+1$ for trajectories on ΓY.

4. Steady-state bifurcations. Let x_0 be in R^n and let $X = \Gamma x_0$. Suppose that $F: R^n \times R \to R^n$ is a smooth family of equivariant vector fields and X is a relative equilibrium of F for all values of λ. Theorem 2.1 guarantees that F can be decomposed as $F = F_N + F_T$, where F_N is a family of normal vector fields and F_T is a family of tangent vector fields. Let $G(\cdot, \lambda)$ denote the restriction of $F_N(\cdot, \lambda)$ to the normal space N_{x_0}. Note that x_0 is an equilibrium of G for all values of λ. We call x_0 the *trivial equilibrium* of G. We say that the family F has a *steady-state bifurcation* near X, if there exists a branch of nontrivial equilibria of G emanating from x_0. Note that such a bifurcation will generically occur if $(dG)_{(x_0,0)}$ has a zero eigenvalue and the action of the isotropy subgroup Σ_{x_0} on the center subspace of $(dG)_{(x_0,0)}$ is nontrivial.

Suppose that F has a steady-state bifurcation. Let $y(\lambda)$, $0 \leqq \lambda < \lambda_0$, be a bifurcating branch of nontrivial equilibria of G. We assume that all the equilibria $y(\lambda)$ have the same isotropy subgroup Σ. We also assume that the map $\lambda \to y(\lambda)$ is smooth on the open interval $(0, \lambda_0)$. Let $Y(\lambda)$ denote the group orbits of the equilibria $y(\lambda)$. Theorem 2.2 guarantees that the sets $Y(\lambda)$ are invariant under the flow of F. The goal of this section is to analyze that flow of F on the sets $Y(\lambda)$.

Let $z(\varepsilon, t)$ be the trajectory of F with initial condition $y(\lambda)$. Equivariance of F implies that each trajectory on $Y(\varepsilon)$ is given as $\gamma z(\lambda, t)$, for some $\gamma \in \Gamma$. Hence, to understand the dynamics on Y it suffices to analyze the structure of $z(\lambda, t)$. Let $N(\Sigma)$ denote the normalizer of Σ. Our analysis is based on the following observations:

(a) The trajectory $z(\lambda, t)$ is contained in $N(\Sigma)y(\lambda)$.

(b) There exists an integer $k \geqq 0$ such that $z(\lambda, t)$ can be described as k-frequency drift along the group orbit Y. More precisely, there exists a k-torus $\mathbb{T} \subset \Gamma$ such that $z(\varepsilon, t)$ is dense in $\mathbb{T}y$.

Field [1980, Prop. B1] has proved that the number of independent frequencies of the drift is bounded by the dimension of a maximal torus in $N(\Sigma)/\Sigma$. The result of Field can be easily deduced from properties (a) and (b).

We now state the main result of this section.

THEOREM 4.1. *For a generic family F the dimension of the drift along the orbit $Y(\lambda)$ equals the dimension of a maximal torus in $N(\Sigma)/\Sigma$ for all except countably many values of λ.*

Theorem 4.1 is an extension of Proposition B1 in Field [1980]. Dancer [1980] also obtained results relevant to the problem discussed in this section. Suppose that f is a smooth Γ-equivariant vector field and y is an equilibrium of f. Let Σ be the isotropy subgroup of y. In the proposition on p. 88 Dancer proved that if $\dim N(\Sigma) > \dim \Sigma$ then generically all equilibria of f which are sufficiently near y lie in the group orbit of y. Property (b) is a generalization of this result.

In the latter part of this section we state a more precise version of Theorem 4.1. In order to do this we need to review some concepts and results from Lie group theory.

Before we can prove Theorem 4.1 we need to analyze the flow on a relative equilibrium of a single vector field f. Let Y be a relative equilibrium of f and suppose that Σ is the isotropy subgroup of some $y \in Y$. We prove that a trajectory on Y is dense in a k-dimensional torus and that generically k equals the dimension of a maximal torus in $N(\Sigma)/\Sigma$.

In Proposition 4.6 we prove an important technical result stating that the set of all images of a point y in R^n under a smooth Γ-equivariant map is Fix (Σ_y). This result is stated without proof in Lemma A of Field [1980].

Proposition 4.10, which is stated following the proof of Theorem 4.1′ describes what happens when the drift fails to be of maximal dimension. The proposition asserts that for a generic family the dimension of the drift can only decrease by 1. Field [1988] proves that if the dimension of a maximal torus in $N(\Sigma)/\Sigma$ equals 1, then for a generic family F the set $Y(\lambda)$ contains no equilibria. This result does not follow from Theorem 4.1.

In order to state a more precise version of Theorem 4.1 we need to review the concepts of maximal tori and rank of a Lie group. Let Δ be a Lie group. We say that a Lie subgroup $\mathbb{T}$ of Δ is a *torus* if $\mathbb{T}$ is compact, Abelian, and connected. A torus is called *maximal* if it is not properly contained in any other torus. The following is the main result on maximal tori.

THEOREM 4.2. *In a Lie group Δ any two maximal tori are conjugate, and every element of Δ is contained in a maximal torus.*

The proof of Theorem 4.2 can be found in Bröcker and tom Dieck [1985, Thm. (1.6), p. 159].

Theorem 4.2 implies that all maximal tori are of the same dimension. The dimension of maximal tori in Δ is called the *rank* of Δ.

Let $l = \operatorname{rank} \Delta$ and let $\xi \in \mathscr{L}(\Delta)$. We say that ξ *generates a maximal torus* in Δ if the set $\{\exp t\xi : t \in R\}$ is dense in a torus of dimension l. We have the following proposition.

PROPOSITION 4.3. *The set of $\xi \in \mathscr{L}(\Delta)$ which generates a maximal torus is residual (an intersection of open and dense sets).*

Proof. Let $\mathbb{T}$ be a maximal torus in Δ. We identify $\mathbb{T}$ with R^l/Z^l and $\mathscr{L}(\mathbb{T})$ with R^l (see Bröcker and tom Dieck [1985, Cor. I, eq. (3.7)]). Let

$$P^m = \{\xi \in \mathscr{L}(\mathbb{T}) : \xi = (\xi_1, \xi_2, \cdots, \xi_l) \text{ and } \sum m_j \xi_j = 0\}$$

and let

$$E^m = \bigcup_{\sigma \in \Delta} \operatorname{Ad}_\sigma P^m.$$

Since the group Δ is compact, it follows that the image of $\mathscr{L}(\Delta)$ under the exponential mapping is the connected component of the identity in Δ (see Bröcker and tom Dieck [1985, Thm. IV, eq. (2.2)]). Hence, by Theorem 4.2, each $\zeta \in \mathscr{L}(\Delta)$ has the form $\operatorname{Ad}_\sigma \xi$,

$\xi \in \mathscr{L}(\mathbb{T})$. Since $\exp \mathrm{Ad}_\sigma \xi = \sigma \exp \xi \sigma^{-1}$ it follows that ζ generates a maximal torus if and only if ξ does. Also ξ generates a maximal torus if it is in the complement of the sets E^m for all $m \in Z^l$. The sets E^m are nowhere dense (see the proof of Theorem 4.1'), so the complement of their union is a residual set.

Throughout we assume the following conditions on the family of vector fields F:

(S1) The orbit X is a trivial relative equilibrium of F. In other words, $F_N(\lambda, x_0) = 0$ for all values of λ.

(S2) There exists $\lambda_0 > 0$ and a branch of relative equilibria of F_N, parametrized as $y(\lambda)$, $0 < \lambda < \lambda_0$. The mapping $\lambda \to y(\lambda)$ is smooth on $(0, \lambda_0)$. The points $y(\lambda)$ have isotropy Σ.

Let $C^\infty_T(R^n \times R, R^n)$ denote the space of smooth families of equivariant tangent vector fields on R^n. For a family F satisfying (S1) and (S2) let $Y(\lambda) = \Gamma y(\lambda)$. We now state Theorem 4.1 more precisely.

THEOREM 4.1'. *Suppose that a family of vector fields F satisfies* (S1) *and* (S2). *Then*

(i) *Trajectories on the manifolds $Y(\lambda)$ are dense in tori of dimension bounded by* $\operatorname{rank}(N(\Sigma)/\Sigma)$.

(ii) *There exists a residual set $\mathscr{B} \subset C^\infty_T(R^n \times R, R^n)$ such that for every $H \in \mathscr{B}$ there exists a countable set $I_0 \subset (0, \lambda_0)$ such that for every $\lambda \in (0, \lambda_0) \setminus I_0$ trajectories of $F + H$ on the manifolds $Y(\lambda)$ are dense in tori of dimension equal to* $\operatorname{rank}(N(\Sigma)/\Sigma)$.

Note that Theorem 4.1' is more general than Theorem 4.1: we assume that $y(\lambda)$ is a branch of relative equilibria of F_N rather than a branch of equilibria of F_N. This assumption does not increase the complexity of the proof.

Before proving Theorem 4.1' we analyze the following simpler situation. Suppose that $g: R^n \to R^n$ is a Γ-equivariant vector field with the following properties:

(VS1) The orbit X is a relative equilibrium of g.

(VS2) There exists $y_0 \in R^n$, $y_0 \notin X$ such that $Y = \Gamma y_0$ is a relative equilibrium of g.

The problem of finding the dynamics on Y has been solved by Field [1980]. Here we briefly present his results. We start with the following proposition.

PROPOSITION 4.4. *Suppose that $g(y_0) = v$. Let $\xi \in \mathscr{L}(\Gamma)$ be such that $\xi y_0 = y$. Then $y(t) = \exp(t\xi) y_0$ is the integral curve of g with $y(0) = y_0$.*

Proof.

$$\begin{aligned}\dot{y}(\tau) &= \frac{d}{dt} \exp(t\xi) y_0 \big|_{t=\tau} \\ &= \exp(\tau\xi) \frac{d}{dt} \exp((t-\tau)\xi) y_0 \big|_{t=\tau} \\ &= \exp(\tau\xi) \frac{d}{ds} \exp(s\xi) y_0 \big|_{s=0}.\end{aligned}$$

By definition $(d/ds) \exp(s\xi) y_0|_{s=0} = \xi y_0$. Hence

$$\begin{aligned}\dot{y}(\tau) &= \exp(\tau\xi) \xi y_0 = \exp(\tau\xi) g(y_0) \\ &= g(\exp(\tau\xi) y_0) = g(y(\tau)).\end{aligned}$$

For $h: R^n \to R^n$ let $\|h\| = \sup_{x \in R^n} |h(x)|$. The following theorem gives a complete description of dynamics on relative equilibria of a vector field g.

THEOREM 4.5 (Field [1980, Prop. B1]). *Suppose that $g: R^n \to R^n$ is an equivariant vector field satisfying* (VS1) *and* (VS2) *and let Σ be the isotropy subgroup of y_0. Then*

(i) *Every flow trajectory contained in Y has the form $z(t)=\exp(t\xi)y$, for some $\xi\in\mathscr{L}(N(\Sigma))$.*

(ii) *The dimension of the torus $\mathbb{T}=\operatorname{cl}\{\exp(t\xi)y_0\colon t\in R\}$ is less than or equal to* $\operatorname{rank}(N(\Sigma)/\Sigma)$.

(iii) *For every $\varepsilon>0$ there exists a vector field h, such that $\|h\|<\varepsilon$, Y is a relative equilibrium of h, and the dimension of the closure of trajectories of $g+h$ on Y equals* $\operatorname{rank}(N(\Sigma)/\Sigma)$.

To prove Theorem 4.5 we need to answer the following question. What are the possible images of the vector y_0 under Γ-equivariant vector fields? Suppose that V is the space of all possible images of y under Γ-equivariant vector fields; that is,

$$V=\{h(y),\ h:R^n\to R^n \text{ is a } \Gamma\text{-equivariant vector field}\}.$$

If h is a Γ-equivariant vector field, then it follows that $h(y_0)$ is fixed by all elements in Σ_{y_0}. Hence $V\subset \operatorname{Fix}(\Sigma_{y_0})$. The following proposition shows that the other containment also occurs.

PROPOSITION 4.6. *The space V is equal to the fixed-point space of Σ_{y_0}; that is,*

$$V=\operatorname{Fix}(\Sigma_{y_0}).$$

Proof. Let $Y=\Gamma y_0$ and let $\Sigma=\Sigma_{y_0}$. Suppose that $v\in\operatorname{Fix}(\Sigma)$. We first show that there exists a smooth and Γ-equivariant vector field $g:R^n\to R^n$ such that $g(y_0)=v$. Let $N(Y)$ be the normal bundle of Y. Recall that $N(Y)$ can be identified with an invariant neighborhood U of Y in R^n. It follows that $TN(Y)$ can be identified with R^n. We define a vector field $h:N_{y_0}\to TN(Y)$ by $h(z)=v$. Clearly, h is smooth and Σ-equivariant. By Proposition 1.1 we can extend h to a smooth, Γ-equivariant vector field g_1 on $N(Y)$. The properties of the normal bundle $N(Y)$ imply that the vector field g_1 can be identified with a vector field g_2 defined on U. Let $\alpha:R^n\to R$ be a smooth, invariant function such that $\alpha(y_0)=1$ and $\alpha(x)=0$ for all $x\notin U$. Let g be defined as follows:

$$g(x)=\begin{cases}\alpha(x)g_2(x) & \text{if } x\in U,\\ 0 & \text{if } x\notin U.\end{cases}$$

Clearly, g is smooth, Γ-equivariant, and $g(y_0)=v$.

Remark 4.7. The vector field h described in Theorem 4.5(iii) can be chosen so that $g+h$ is a polynomial vector field. The existence of such h can be shown using the equivariant version of the Stone-Weierstrass approximation theorem (see Poenaru [1976, proof of Prop. 1, p. 20]).

The final ingredient necessary to prove Theorem 4.5 is given by the following elementary lemma.

LEMMA 4.8. *The following equality holds for any $y\in R^n$:*

$$T_yY\cap\operatorname{Fix}(\Sigma_y)=\{\xi y\colon \xi\in\mathscr{L}(N(\Sigma_y))\}.$$

Proof. Let $\xi\in\mathscr{L}(\Gamma)$, $y\in R^n$. Then $\xi y\in\operatorname{Fix}(\Sigma_y)$ if and only if $\exp\xi y\in\operatorname{Fix}(\Sigma_y)$. This implies that $\sigma\exp\xi y=\exp\xi y$ for all $\sigma\in\Sigma_y$. It follows that there exists $\eta\in\mathscr{L}(N(\Sigma))$ such that $\eta y=\xi y$. The lemma now follows.

Proof of Theorem 4.5. The theorem is an easy consequence of Proposition 4.4, Proposition 4.6, and Lemma 4.8.

In the remainder of this section we prove Theorem 4.1′. Let F be a family of vector fields satisfying (S1) and (S2). We begin by defining a map which assigns to each $H\in C^\infty_\Gamma(R^n\times R, R^n)$ a curve ξ in $\mathscr{L}(N(\Sigma)/\Sigma)$ such that $(F+H)(y(\lambda),\lambda)=\xi(\lambda)y(\lambda)$ for each λ in some interval I. Recall that $T_yY\cap\operatorname{Fix}\Sigma=\{\xi y\colon \xi\in\mathscr{L}(N(\Sigma))\}$.

Let V_Σ be the set of $y \in R^n$ with isotropy group Σ. For $y \in V_\Sigma$ the space $\{\xi y\colon \xi \in \mathscr{L}(N(\Sigma))\}$ is isomorphic to $\mathscr{L}(N(\Sigma)/\Sigma)$. Let $\Xi\colon T_y Y \cap \operatorname{Fix}\Sigma \to \mathscr{L}(N(\Sigma)/\Sigma)$ denote this isomorphism. It is clear that Ξ changes smoothly as y is being varied in V_Σ. Let I be a subinterval of $(0, \lambda_0)$. The map $\Theta\colon C_T^\infty(R^n \times R, R^n) \to C^\infty(I, \mathscr{L}(N(\Sigma)/\Sigma))$ is defined as

$$\Theta(H)(\lambda) = \Xi((F+H)(y(\lambda), \lambda)).$$

We prove Theorem 4.1′ by showing that for a residual set of $H \in C_T^\infty(R^n \times R, R^n)$ the curve $\Theta(H)$ is transverse to all the sets E^m (see Proposition 4.3). Before presenting the proof we review some concepts related to the Whitney C^∞ topology. For a more complete treatment of this topic see Golubitsky and Guillemin [1974]. Let Z, W be smooth manifolds. For a positive integer q let $J^q(Z, W)$ denote the space of q-jets of smooth maps from Z to W. We describe a neighborhood basis of a map f in the Whitney C^∞ topology on $C^\infty(Z, W)$. Let q be a positive integer and let d_q be a metric on $J^q(Z, W)$ compatible with its topology (such a metric exists by (I, 5.9) in Golubitsky and Guillemin [1974]). Let $\delta\colon Z \to R^+$ be a continuous function. Let

$$U_{q,\delta} = \{g \in C^\infty(Z, W)\colon d_q(j^q f(x), j^q g(x)) < \delta(x) \text{ for all } x \in Z\}.$$

The collection of the sets $U_{q,\delta}$ for all choices of q and δ forms a neighborhood basis of f in the Whitney C^∞ topology.

Suppose that Z is an open subset of R^p for some p, and W is a vector space. Then the above-mentioned metrics d_q can be chosen as follows. Suppose $s = \dim W$. We identify W with R^s. For a positive integer q and $g \in J^q(Z, W)$ let

$$\|g\|_q(x) + |x| + |g(x)| + \sum_{1 \le |\alpha| \le q} \left| \frac{\partial^{|\alpha|} g(x)}{\partial x^\alpha} \right|.$$

Here α denotes a p-vector of nonnegative integers. We define d_q on $J^q(Z, W)$ as

$$d_q(\sigma_1, \sigma_2) = \|g_1 - g_2\|_q(x),$$

where $\sigma_1, \sigma_2 \in J^q(Z, W)$ and g_1, g_2 are such that $\sigma_1 = j^q g_1(x)$ and $\sigma_2 = j^q g_2(x)$. It is easy to see that d_q agrees with the topology on $J^q(Z, W)$.

Let I be the interval used in the definition of the map Θ. Then we have Lemma 4.9.

LEMMA 4.9. *If $\bar{I} \subset (0, \lambda_0)$ then Θ is continuous in the Whitney C^∞ topology.*

Proof. Let $C(I) = \{(y(\lambda), \lambda)\colon \lambda \in I\}$. In this proof we use the metrics d_q described prior to the statement of Lemma 4.9. The map Θ can be written as

$$\Theta(H) = \Xi \circ H|C(I).$$

Hence Θ is a composition of two maps: a map Θ_1 given as

$$\Theta(H) = H|C(I)$$

and a map Θ_2 defined as

$$\hat{H} \to \Xi \circ \hat{H}.$$

The map Θ_1 does not, in general, have to be continuous, but it is continuous if $\bar{I} \subset (0, \lambda_0)$. This follows, since $\bar{I} \subset (0, \lambda_0)$ implies that for any given q all the partial derivatives of the function $\lambda \mapsto y(\lambda)$ are bounded on I. Hence continuity of Θ_1 can be established through repeated application of the chain rule. The map Θ_2 is continuous by (II, 3.5) in Golubitsky and Guillemin [1974].

Proof of Theorem 4.1′. Part (i) of the theorem follows from Theorem 4.5.

We now prove part (ii). Suppose that l is the rank of $N(\Sigma)/\Sigma$. Let $\mathbb{T}$ be a maximal torus in $N(\Sigma)/\Sigma$. As in the proof of Proposition 4.3 we identify $\mathbb{T}$ with R^l/Z^l and $\mathcal{L}(\mathbb{T})$ with R^l. Recall the definitions of the sets E^m and P^m (see the proof of Proposition 4.3). In the proof of Proposition 4.3 we show that the sets E^m have the following property: a vector $\xi \in \mathcal{L}(N(\Sigma)/\Sigma)$ generates a maximal torus if and only if ξ is in the complement of E^m for all m in Z^l.

The sets E^m may not be manifolds, but we show that each E^m is a finite union of manifolds. Let $\xi \in P^m$ and let Δ be the isotropy subgroup of ξ with respect to the adjoint action of $N(\Sigma)/\Sigma$ on $\mathcal{L}(N(\Sigma)/\Sigma)$. Let $[\cdot, \cdot]$ denote the bracket in $\mathcal{L}(N(\Sigma)/\Sigma)$. It is known (see Bröcker and tom Dieck [1985, I, eq. (2.12)]) that

$$[\eta, \zeta] = \frac{d}{dt} \operatorname{Ad}_{\exp t\eta} \zeta|_{t=0} \quad \text{for all } \eta, \zeta \in \mathcal{L}(N(\Sigma)/\Sigma). \tag{4.1}$$

The containment $P^m \subset \mathcal{L}(\mathbb{T})$ implies that $[\xi, \eta] = 0$ for all $\eta \in P^m$. Let $O(\xi) = \bigcup_{\sigma \in N(\Sigma)/\Sigma} \operatorname{Ad}_\sigma \xi$ be the orbit of ξ. Let U be a small neighborhood of $e\Delta$ in $(N(\Sigma)/\Sigma)/\Delta$ and let $\sigma: U \to N(\Sigma)/\Sigma$ be a local cross section. Equation (4.1) implies that $T_\xi O(\xi) \cap P^m = \{0\}$. It follows that the map $\Psi: U \times P^m \to \mathcal{L}(N(\Sigma)/\Sigma)$, given by $\Psi(u, \eta) = \operatorname{Ad}_{\sigma(u)} \eta$, is a local diffeomorphism near (e, ξ).

Let $E^m(\Delta)$ be the set of all elements of E^m whose isotropy subgroup (with respect to the adjoint action) is conjugate to Δ, and let $P^m(\Delta) = P^m \cap E^m(\Delta)$. Note that $P^m(\Delta)$ is an open subset of $P^m \cap \operatorname{Fix}(\Delta)$. For every $\xi \in P^m$ the corresponding map Ψ is a local diffeomorphism near (e, ξ). It follows that $E^m(\Delta)$ is a smooth manifold.

Note that the number of the sets $E^m(\Delta)$ is finite. This follows from the fact that $N(\Sigma)/\Sigma$, being a compact group, has a finite number of conjugacy classes of isotropy subgroups. Clearly, $E^m = \bigcup E^m(\Delta)$.

The theorem follows from the following assertion:

(*) For every $m \in Z^l$, $\mu \in (0, \lambda_0)$, there exists an interval I containing μ and a set $\mathcal{B}^m(I) \subset C^\infty_T(R^n \times R, R^n)$ with the following properties:
(1) $\mathcal{B}^m(I)$ is residual in the C^∞ Whitney topology.
(2) $\Theta(H)$ is transverse to all the sets $E^m(\Delta)$ at each $\lambda \in I$.

We first show that the theorem follows from (*). To see this let $I_1^m, I_2^m, \cdots$ be a sequence of intervals such that $\bigcup_{i=1}^\infty I_i^m = (0, \lambda_0)$ and $\mathcal{B}^m(I_i^m)$ satisfies the properties (1) and (2). Let $\mathcal{B} = \bigcap_{i=1,j=1}^\infty \mathcal{B}^j(I_i^j)$. It follows that for every $H \in \mathcal{B}$ the curve $\Theta(H)$ is transverse to all the sets E^m at every $\lambda \in (0, \lambda_0)$. It is clear that $\mathcal{B}$ satisfies the property required in the statement of Theorem 4.1′.

We now prove (*). Fix $m \in Z^l$, a subgroup $\Delta \subset N(\Sigma)/\Sigma$ and $\mu \in (0, \lambda_0)$. Let $I \subset I_0 \subset (0, \lambda_0)$ be intervals with $\mu \in I$ and $\bar{I}_0 \subset (0, \lambda_0)$. Let $\mathcal{A}_0$ be the set of $\xi \in C^\infty(I_0, \mathcal{L}(N(\Sigma)/\Sigma))$ such that ξ is transverse to $E^m(\Delta)$ at each $\lambda \in I$. By standard transversality arguments (see Golubitsky and Guillemin [1974, (II, 4.5)]) the set $\mathcal{A}_0$ is open and dense in the Whitney C^∞ topology. We assume that I_0 is the interval used in the definition of the map Θ. Let $\mathcal{A} = \Theta^{-1}\mathcal{A}_0$. It follows from Lemma 4.9 that $\mathcal{A}$ is an intersection of open sets. We now show that $\mathcal{A}$ is dense. Fix $H \in C^\infty_T(R^n \times R, R^n)$. We construct a sequence of families $\{H_i\}$ converging to H and such that each $H_i \in \mathcal{A}$. Let $\{\xi_i\}$ be a sequence of elements of $C^\infty(I_0, \mathcal{L}(N(\Sigma)/\Sigma))$ such that each curve $\Theta(H) + \xi_i$ is in $\mathcal{A}_0$ and the curves ξ_i converge to the zero curve as $i \to \infty$. Such a sequence exists, since $\mathcal{A}_0$ is dense in $C^\infty(I_0, \mathcal{L}(N(\Sigma)/\Sigma))$. To show the existence of the sequence $\{H_i\}$ it suffices to prove that for every $\eta \in C^\infty(I_0, \mathcal{L}(N(\Sigma)/\Sigma))$ there exists a family H_η such that $\Theta(H_\eta)(\lambda) = \eta(\lambda)$ for all $\lambda \in I$ and such that for every positive integer q the size of partial derivatives of H_η of order less than or equal to q can be estimated by the

size of partial derivatives of η of order less than or equal to q. We now give a more precise description of this estimate. Let $\hat{\eta}$ be a smooth curve of elements of $\mathscr{L}(N(\Sigma))$ which projects to η in $\mathscr{L}(N(\Sigma)/\Sigma)$. Suppose that α is an n-vector of positive integers and β a positive integer. Then there exists a constant C, depending only on m, and such that

$$\left|\frac{\partial^{|\alpha|+\beta} H_\eta(z,\lambda)}{\partial z^\alpha \partial \lambda^\beta}\right| \leqq C \left|\frac{\partial^\beta \hat{\eta}(\lambda)}{\partial \lambda^\beta}\right| \tag{4.2}$$

for all $(z,\lambda) \in V \times I_0$. Moreover, $H_\eta(z,\lambda) = 0$ for $(z,\lambda) \notin V \times I_0$.

We now construct H_η. Let $y_0 = y(\mu)$ and $Y = \Gamma y_0$. Recall that $N(Y)$ is equivariantly diffeomorphic to an invariant neighborhood of Y in R^n. Let V be such a neighborhood. In the sequel we identify V with $N(Y)$. By shrinking the interval I we can assume that $y(\lambda) \in V$ for all $\lambda \in I$. We can assume that $y(\lambda) \in N_{y_0}$ for all $\lambda \in I$. Otherwise, we could replace the curve $y(\lambda)$ by a curve $\hat{y}(\lambda) = \gamma(\lambda) y(\lambda)$ with $\gamma(\lambda) \in N(\Sigma)/\Sigma$. We can now define $\hat{H}(z,\lambda)$ as $\hat{\eta}(\lambda) z$ is $z \in N_{y_0}$ and extend this definition by equivariance (see the proof of Proposition 4.6). Let U_0 be a small neighborhood of $e\Sigma$ in Γ/Σ and let σ be a local cross section of π (see § 1). Let $\phi : U_0 \times N^\varepsilon_{y_0}$ be defined as $\phi(u, y) = \sigma(u) y$ (here $N^\varepsilon_{y_0}$ denotes a disc of radius ε around y_0 in N_{y_0}). Recall that for ε small enough ϕ is a diffeomorphism. Let $U = \phi(U_0 \times N^\varepsilon_{y_0})$. We can express every point $z \in U_0$ in local coordinates as $\sigma(u) y$, $y \in N_{y_0}$, $u \in U$. Then, for every $z \in U$, $\hat{H}(z,\lambda) = \sigma(u)\hat{\eta}(\lambda) z$. It is clear from this expression and from the smoothness of the action of Γ that (4.2) holds for all $z \in U$. From compactness of Y it follows that the bound (4.2) holds on a neighborhood V_1 of Y in R^n with possibly a different constant C. With no loss of generality we can assume that $V = V_1$. Let W be an invariant neighborhood of y_0 such that $\bar{W} \subset V$ and suppose that I is chosen so that $y(\lambda) \in W$ for all $\lambda \in I$. Let $h : R^n \times R \to R$ be a smooth Γ-invariant cutoff function vanishing on the complement of $V \times I_0$ and equal to 1 on $U \times I$. Let $H_\eta(z,\lambda) = h(z,\lambda)\hat{H}(z,\lambda)$. It is clear that H_η is globally defined and satisfies (4.2).

We complete the proof of (*) by defining $\mathscr{B}^m(I)$ as the intersection of the sets $\mathscr{A}$ for all choices of Δ.

For $m_1, m_2 \in Z^l$ let $E^{m_1,m_2} = E^{m_1} \cap E^{m_2}$. Note that if m_1 and m_2 are not collinear then the sets E^{m_1,m_2} have codimension 2 in $\mathscr{L}(N(\Sigma)/\Sigma)$. The union of these sets consists of the elements $\zeta \in \mathscr{L}(N(\Sigma)/\Sigma)$ which generate a torus of dimension no less than rank $(N(\Sigma)/\Sigma) - 1$. In the proof of Theorem 4.1′ we could, instead of the sets E^m, use the sets E^{m_1,m_2}. Then, for every $H \in \mathscr{B}$, the curve $\Theta(H)$ would be transverse to all E^{m_1,m_2} at each $\lambda \in (0, \lambda_0)$. This would imply that if m_1 and m_2 were not collinear then E^{m_1,m_2} and $\Theta(H)$ would not intersect. This property implies the following proposition.

PROPOSITION 4.10. *Suppose that F satisfies* (S1) *and* (S2). *Then there exists a residual set $B \subset C^\infty_T(R^n \times R, R^n)$ such that if $H \in B$ then the dimension of the trajectories of $F + H$ on the sets Y is greater than or equal to* rank $(N(\Sigma)/\Sigma) - 1$.

5. Hopf bifurcations. Let x_0 be in R^n and let $X = \Gamma x_0$. Suppose that $F : R^n \times R \to R^n$ is a smooth family of equivariant vector fields and X is a relative equilibrium of F for all values of λ. By Theorem 2.1 $F = F_N + F_T$, where F_N is a family of normal vector fields and F_T is a family of tangent vector fields. Let G be the family defined at the beginning of § 4; that is, $G(\cdot,\lambda)$ is the restriction of $F_N(\cdot,\lambda)$ to the normal space N_{x_0}. Recall that x_0 is the trivial equilibrium of G. We say that the family F has a *Hopf bifurcation* near X, if there exists a branch of nontrivial periodic orbits of G emanating from x_0. Note that such a bifurcation will generically occur if $(dG)_{(x_0,0)}$ has a purely imaginary eigenvalue.

Suppose that F has a Hopf bifurcation. Let $Y(\lambda)$, $\lambda_0 > \lambda > 0$, be a branch of periodic orbits of G and let $y(\lambda)$ denote the initial conditions for the trajectories $Y(\lambda)$. We assume that all the points $y(\lambda)$ have the same isotropy subgroup Σ. Let $\Sigma_{Y(\lambda)}$ be the *group of symmetries* of the set $Y(\lambda)$; that is,

$$\Sigma_{Y(\lambda)} = \{\sigma \in \Sigma_{x_0} : \sigma Y(\lambda) = Y(\lambda)\}.$$

We assume that all the sets $Y(\lambda)$ have the same group of symmetries Σ_Y.

Let $Z(\lambda)$ denote the group orbits of the sets $Y(\lambda)$. Theorem 2.2 guarantees that the sets $Z(\lambda)$ are invariant under the flow of F. The goal of this section is to analyze that flow of F on the sets $Z(\lambda)$. Our analysis is based on the following result: every trajectory of F on Z is dense in a $(k+1)$-dimensional torus, with k frequencies given by the drift along group orbits and the additional frequency corresponding to the motion along the periodic orbit Y. This result was obtained by Field [1980, Prop. B2]. Field also showed that the number of the drift frequencies is bounded by rank $(N(\Sigma)/\Sigma)$.

This section contains two main results. The first of these results is a modification of the theorem of Field. We assume that f is a smooth Γ-equivariant vector field, X is a relative equilibrium of f, and Y is a periodic orbit of f_N. Let $Z = \Gamma Y$. The theorem states that the trajectories on Z are dense in tori of dimension $k+1$, with $k \leqq$ rank $N(\Sigma_Y)/\Sigma_Y$. For some choices of f rank $(N(\Sigma_Y)/\Sigma_Y) <$ rank $(N(\Sigma)/\Sigma)$. This is illustrated in Example 5.3.

Our second main result (Theorem 5.2) deals with the dynamics of the family of vector fields F on the sets $Z(\lambda)$. The theorem states that, given a generic family of vector fields F, there exists a countable set $L_0 \subset (0, \lambda_0)$ such that if $\lambda \notin L_0$ then the trajectories on $Z(\lambda)$ are dense in tori of maximal dimension. In Proposition 5.7 we strengthen this result by showing that generically the dimension of trajectories on $Z(\lambda)$ drops only by 1.

We now state the first of the two main theorems. Let $f: R^n \to R^n$ be a smooth, Γ-equivariant vector field with the following properties:

(VH1) The orbit X is a relative equilibrium of f.

(VH2) The vector field f_N has a periodic orbit $Y = \{y(t) : t \in [0, T]\}$, where T is the period of Y.

Let Σ_Y denote the group of symmetries of Y and let $Z = \Gamma Y$. We have Theorem 5.1.

THEOREM 5.1. *All trajectories on the set Z are dense in $(k+1)$-dimensional tori, where $k \leqq$ rank $N(\Sigma_Y)/\Sigma_Y$. For every $\varepsilon > 0$ there exists a smooth and Γ-equivariant vector field h such that $\|h\| \leqq \varepsilon$, h satisfies* (VH1) *and* (VH2), *and such that the trajectories of $f+h$ on the set Z are dense in tori of dimension equal to* rank $N(\Sigma_Y)/\Sigma_Y + 1$.

We now state the second main theorem. Let $F: R^n \to R^n$ be a smooth family of Γ-equivariant vector fields with the following properties:

(H1) The orbit X is a relative equilibrium of F for all values of $\lambda \in R$.

(H2) There exists $\lambda_0 > 0$ and a branch of periodic orbits of F_N, parametrized as $(\lambda, Y(\lambda))$, $0 < \lambda < \lambda_0$, with initial conditions y_λ. All the elements y_λ have the same isotropy subgroups Σ and all the sets $Y(\lambda)$ have the same group of symmetries Σ_Y. The map $\lambda \mapsto y_\lambda$ is smooth on the interval $(0, \lambda_0)$.

For a family of vector fields satisfying (H1) and (H2), let $Z(\lambda) = \Gamma Y(\lambda)$. We have the following theorem.

THEOREM 5.2. *If F is a family of vector fields satisfying* (H1) *and* (H2) *then there exists a set $\mathcal{B} \subset C_T^\infty(R^n \times R, R^n)$ with the following properties*:

(i) *For every $H \subset \mathcal{B}$ there exists a countable set $I_0 \subset (0, \lambda_0)$ such that for every $\lambda \in (0, \lambda_0) \setminus I_0$ trajectories of $F+H$ on the manifolds $Z(\lambda)$ are dense in tori of maximal dimension.*

(ii) *The set $\mathcal{B}$ is residual in the Whitney C^∞ topology.*

We now present an example of a vector field f for which rank $(N(\Sigma)/\Sigma) >$ rank $N(\Sigma_Y)/\Sigma_Y$.

Example 5.3. Let $F: R^n \times R \to R^n$ be a smooth family of vector fields equivariant under the action of $O(2)$. Let $\kappa \in O(2) \setminus SO(2)$. Suppose that the invariant equilibrium $x = 0$ bifurcates to a branch of equilibria $x(\lambda)$ with $\Sigma_{x(\lambda)} = \{e, \kappa\}$. Consider a secondary Hopf bifurcation occurring along the branch $x(\lambda)$. In other words, suppose that $(dF_N)_{(x(\lambda_0),\lambda_0)}$ has, for some λ_0, a pair of purely imaginary eigenvalues $i\omega$. Let V be the real eigenspace of $i\omega$. We assume that V is two-dimensional and that the action of κ on V is nontrivial. Then, by the standard Hopf bifurcation theorem, F_N has a branch of periodic solutions with period T and such that $y_\lambda(t + (T/2)) = \kappa y_\lambda(t)$. By Theorem 2.2 the trajectory of F corresponding to $y_\lambda(t)$ is $z_\lambda(t) = \gamma(t) y_\lambda(t)$, where $\gamma(t) \in SO(2)$ and $\gamma(0) = e$. Let Φ_t denote the flow of F. Then

$$x_\lambda(t) = \Phi_T y_\lambda(0) = \Phi_{T/2}(\Phi_{T/2} y_\lambda(0)) = \Phi_{T/2}(\gamma(T/2)\kappa y_\lambda(0)).$$

Equivariance of $\Phi_{T/2}$ implies that

$$\Phi_{T/2}(\gamma(T/2)\kappa y_\lambda(0)) = (\gamma(T/2)\kappa)^2 y_\lambda(0) = y_\lambda(0) = x_\lambda(0).$$

It follows that $x_\lambda(t)$ must be a periodic solution. Note that $\Sigma = \{e\}$, so $N(\Sigma)/\Sigma = O(2)$, but $\Sigma_Y = \{e, \kappa\}$, so $N(\Sigma_Y)/\Sigma_Y$ is discrete. Hence rank $(N(\Sigma)/\Sigma) >$ rank $(N(\Sigma_Y)/\Sigma_Y)$.

In the remainder of the section we prove Theorems 5.1 and 5.2. In the proof of Theorem 5.1 we will use two lemmas and some background information from Lie group theory. We begin by stating and proving the first of the lemmas. Note that $\Sigma \subset \Sigma_Y \subset N(\Sigma)$. Let $\Delta = \Sigma_Y/\Sigma$. We have the following lemma.

LEMMA 5.3. *The group Δ is finite and cyclic or Δ is isomorphic to S^1.*

Proof. Let y_0 be the initial condition of the periodic orbit Y. We assume, with no loss of generality, that the period of the solution $y(t)$ is 1. We identify S^1 with R/Z. Note that Y is diffeomorphic to S^1 via the map $t \to y(t)$. Let $\rho: \Delta \to S^1$ be defined by the identity

$$y(\rho(\delta)) = \delta y_0, \qquad \delta \in \Delta.$$

Clearly, ρ is smooth and well defined. Note that the action of Δ on Y is free, so ρ must be injective. Equivariance of f implies that ρ is a Lie group homomorphism. It follows that $\rho(\Delta)$ (which is isomorphic to Δ) is a Lie subgroup of S^1 and therefore must be isomorphic to either S^1 or Z_l for some l.

We now review some of the concepts from Lie group theory, which will be used in the proof of Theorem 5.1. For a Lie group H let H_0 denote the connected component of the identity in H. We say that a subgroup K of a compact Lie group H is *topologically cyclic* if there exists $\delta \in K$ such that $K = \text{cl}\,\{\delta^n : n \text{ is an integer}\}$. The element δ is called the *generator* of K. We say that K is a *Cartan subgroup* of H if K is topologically cyclic and $N(K)/K$ is discrete. In the proof of Theorem 5.1 we will use the following two propositions.

PROPOSITION 5.4. *Each element $h \in H$ is contained in a Cartan subgroup K of H such that K/K_0 is generated by hK_0.*

PROPOSITION 5.5. *If K is a Cartan subgroup of H generated by z, then any $h \in H_0 z$ is conjugate to an element of $K_0 z$ via conjugation by an element of H_0.*

The statements and the proofs of Propositions 5.4 and 5.5 can be found in Bröcker and tom Dieck [1985, IV, eqs. (4.2) and (4.3)].

If K is a Cartan subgroup and K_1 a topologically cyclic group, then K/K_1 must be discrete. Proposition 5.4 implies that each topologically cyclic group is contained in a Cartan subgroup. It follows that a Cartan subgroup can be defined as a topologically cyclic group of maximal dimension.

Let K be a topologically cyclic subgroup of H. Then K_0 must be a torus in H_0 and K/K_0 must be finite and cyclic. It follows that K is isomorphic to $K_0 \times Z_l$, for some l (see Bröcker and tom Dieck [1985, I, eq. (4.14)]).

Let f be a smooth Γ-equivariant vector field satisfying (VH1) and (VH2). Theorem 2.2 guarantees that the trajectory of f with initial condition $y(0)$ has the form $\gamma(t)y(t)$, where $\gamma(t) \in (N(\Sigma)/\Sigma)_0$. We assume that Δ is finite and cyclic with generator δ_0 (see Lemma 5.3) and let T_0 be defined by the identity $y(T_0) = \delta_0 y_0$. To prove Theorem 5.1 we need to prove the following lemma.

LEMMA 5.6. *Suppose that $\gamma(T_0) = \gamma_0$ and that $\gamma_1 \in (N(\Sigma)/\Sigma)_0$. Then there exist a smooth, Γ-equivariant tangent vector field h and a curve $\gamma^1(t)$ such that:*

(i) $\gamma^1(0) = e$.

(ii) $\gamma^1(T_0) = \gamma_1$.

(iii) *$\gamma^1(t)y(t)$ is the trajectory of $f+h$.*

Proof. Let $\xi(t) = \gamma(t)^{-1}\dot{\gamma}(t)$. Note that $f_T(y(t)) = \xi(t)y(t)$. Let $\sigma = \gamma_1\gamma_0^{-1}$ and let $\zeta = \exp^{-1}\sigma$. Let $\eta(t)$ be a smooth function such that $\eta(0) = 0$, $\eta(T_0) = 1$, and

$$\frac{d^j}{dt^j}\eta(0) = \frac{d^j}{dt^j}\eta(T_0) = 0, \qquad j = 1, 2, \cdots.$$

Let $\gamma^1(t) = \exp(\eta(t)\zeta)\gamma(t)$. We define a curve $\hat{\xi}(t) = \xi(t) + \dot{\eta}(t)\gamma(t)^{-1}\zeta\gamma(t)$. Let $\xi^1(t) = \hat{\xi}(t) - \xi(t)$. Note that

$$\frac{d^j}{dt^j}\xi^1(0) = \frac{d^j}{dt^j}\xi^1(T_0) = 0, \qquad j = 1, 2, \cdots. \tag{5.1}$$

We extend the definition of ξ^1 to all of R by requiring that $\xi^1(t + T_0) = \mathrm{Ad}_\delta \xi^1(t)$, where δ is as defined in Lemma 5.3. Equation (5.1) implies that this extension is smooth. Let $h(y(t)) = \xi^1(t)y(t)$. Let $Z = \Gamma Y$. We extend h to Z by equivariance. Let $N(Z)$ be the normal bundle of Z. Recall that $N(Z)$ can be identified with a Γ-invariant neighborhood of Z. We extend h to $N(Z)$ by letting $h(w) = \tilde{h}(z)$ for $w \in N_z$. We use an invariant cutoff function to extend the definition of h to all of R^n. A simple computation shows that $\gamma^1(t)y(t)$ is a trajectory of $f+h$. It is also clear that $\|h\| \to 0$ as σ approaches e.

Proof of Theorem 5.1. Let Δ be as defined prior to the statement of Lemma 5.3; that is, $\Delta = \Sigma_Y/\Sigma$. Lemma 5.3 implies that Δ is either finite and cyclic or isomorphic to S^1. We divide the analysis in two cases:

(1) Δ is isomorphic to S^1.

(2) Δ is finite and cyclic.

Case (1). It follows from the proof of Lemma 5.3 that Z is a relative equilibrium of f. Therefore the dynamics on Z is described by Theorem 4.5. Hence the trajectories on Z are dense in tori of maximal dimension equal to rank $N(\Sigma)/\Sigma$. We now show that rank $N(\Sigma)/\Sigma$ = rank $N(\Sigma_Y)/\Sigma_Y + 1$. Note that Δ is a torus of dimension 1 contained in $N(\Sigma)/\Sigma$. Hence there is a maximal torus $\mathbb{T}$ in $N(\Sigma)/\Sigma$ such that $\Delta \subset \mathbb{T}$. Clearly, $\mathbb{T} \subset N(\Sigma_Y)/\Sigma$. It follows that $\mathbb{T}/\Delta$ is a maximal torus in $N(\Sigma_Y)/\Sigma_Y$, which proves the required equality.

Case (2). Let $z(t)$ be a trajectory on Z. By Theorem 2.2 $z(t)=\gamma(t)y(t)$, $\gamma(t)\in\Gamma$. Since $\Sigma_{y(t)}=\Sigma$ we can assume that $\gamma(t)\in N(\Sigma)/\Sigma$. Let T_0 be the number defined prior to the statement of Lemma 5.6 and let $\gamma_0=\gamma(T_0)\delta_0$. Clearly, $z(T_0)=\gamma_0 y_0$. Let $\mathbb{T}=\operatorname{cl}\{z(t): t\in R\}$ and let

$$S=\operatorname{cl}\{\gamma_0^k: k \text{ is an integer}\}.$$

Observe that $Sy_0\subset\mathbb{T}$. Let Φ_t be the flow of f. We have

$$\Phi_t\gamma_0 y_0=\Phi_t\Phi_{T_0}y_0=\Phi_{t+T_0}y_0=z(t+T_0).$$

It follows that $\Phi_t Sy_0\subset Z$ for all $t\in R$. Let us define the action of $\Gamma\times R$ on R^n as

$$(t,\gamma)w=\Phi_t\gamma w \quad \text{where } (t,\gamma)\in\Gamma\times R,\quad w\in R^n. \tag{5.2}$$

It follows that $\mathbb{T}=(S\times R)y_0$. Let Δ_0 be the isotropy subgroup of y_0 with respect to the action defined by (5.2). It is clear that $(S\times R)/\Delta_0$ is compact, connected, and Abelian. Hence $(S\times R)/\Delta_0$ is isomorphic to a torus. This implies that $\mathbb{T}$ is diffeomorphic to a torus. Note that S is topologically cyclic and γ_0 is its generator. Hence the dimension of S is less than or equal to the dimension of a Cartan subgroup containing γ_0. Note that γ_0 and δ_0 lie in the same component of the identity in $N(\Sigma)/\Sigma$, which we denote by $S(\delta_0)$. Let K be a Cartan subgroup generated by an element $\gamma_1\in S(\delta_0)$. Proposition 5.5 implies that $\dim S\leqq\dim K$ and $\dim\mathbb{T}\leqq\dim K+1$.

We now prove that for a small perturbation of f the dimension of the closure of a trajectory on Z equals $\dim K+1$. Suppose that $\gamma_1\in S(\delta_0)$ generates a Cartan subgroup. By Lemma 5.6 there exists a vector field h satisfying (VH1) and (VH2) and such that if $\tilde{z}(t)=\tilde{\gamma}(t)y(t)$ is the trajectory of $f+h$ with initial condition y_0 then $\tilde{\gamma}(T_0)\delta_0=\gamma_1$. Clearly, the dimension of the closure of $\tilde{z}(t)$ equals $\dim K+1$.

To conclude the proof we need to show that $\dim K=\operatorname{rank} N(\Sigma_Y)/\Sigma_Y$. By Proposition 5.4 we can choose K so that $\Delta\subset K$. The definition of Δ and the fact that Δ is discrete imply that $\operatorname{rank}(N(\Sigma_Y)/\Sigma)=\operatorname{rank}(N(\Sigma_Y)/\Sigma_Y)$. Let $N(\Delta)$ denote the normalizer of Δ in $N(\Sigma)/\Sigma$. Note that $N(\Delta)=N(\Sigma_Y)/\Sigma$. Clearly, $K\subset N(\Delta)$. We show that $\operatorname{rank} N(\Delta)=\dim K$. Suppose that $\phi\in N(\Delta)_0$, and let $\mathbb{T}_0$ be the torus generated by ϕ. Let $\phi_0=\phi\delta_0$ and let $\hat{K}$ be the topologically cyclic subgroup generated by ϕ_0. Since $\phi\in N(\Delta)$ we have $\phi\Delta\phi^{-1}=\Delta$, which implies that

$$\phi\delta_0=\delta_0^m\phi \quad \text{for some } m. \tag{5.3}$$

Note that continuity implies that m is independent of ϕ. It follows that for any positive integer j we must have $\phi_0^j=\delta_0^s\phi^j$ for some s (depending on j but independent of ϕ). Since ϕ_0 is the generator of $\hat{K}$ it follows that for some l we must have $\phi_0^l\in\hat{K}_0\subset N(\Delta)_0$. By continuity we must have $\phi_0^l\in N(\Delta)_0$ for all $\phi\in N(\Delta)_0$. Now (5.3) implies that $\delta_0^s\phi^l\in N(\Delta)_0$ for some s (independent of $\phi\in N(\Delta)_0$). It follows that for some ϕ the torus generated by $\delta_0^s\phi^l$ is a maximal torus in $N(\Delta)$. It follows that $\dim\mathbb{T}_0\leqq\dim\hat{K}\leqq\dim K$. The inequality $\dim K\geqq\operatorname{rank} N(\Delta)$ follows from the fact that K_0 is a connected Abelian subgroup of a compact Lie group; hence it is contained in a maximal torus. It follows that $\operatorname{rank} N(\Delta)\geqq\dim K$.

Let F be a family of vector fields satisfying (H1) and (H2). Suppose that Δ is finite and cyclic and let T_0 be as defined prior to the statement of Lemma 5.6. Let $x_\lambda(t)$ be the trajectory of $F(\cdot,\lambda)$ with initial condition y_λ. By Theorem 2.2 $x_\lambda(t)=\gamma_\lambda(t)y_\lambda(t)$ ($y_\lambda(t)$ is the periodic orbit of $F_N(\cdot,\lambda)$). Let $\gamma(\lambda)=\gamma_\lambda(T_0)$. The proof of Theorem 5.2 is based on the following assertion: a generic F_T gives rise to a generic curve γ. Given H satisfying (H1) and (H2) let $\gamma_\lambda^H(t)y_\lambda(t)$ be the trajectory of $H(\cdot,\lambda)$.

We now define a map Θ which assigns to every family of vector fields satisfying (H1) and (H2) the corresponding curve $\gamma(\lambda)$. Let $I \subset (0, \lambda_0)$ be an open interval and let

$$\Theta(H)(\lambda) = \gamma_\lambda^H(T_0), \qquad \lambda \in I.$$

Let $S(\delta_0)$ be the connected component of $N(\Sigma)/\Sigma$ containing δ_0. Note that $\Theta(H)$ is an element of $C^\infty(I, S(\delta_0))$.

Proof of Theorem 5.2. The proof is analogous to the proof of Theorem 4.1′(ii). We will therefore only outline the proof and omit the technical details. If Δ is isomorphic to S^1 then F satisfies the assumptions of Theorem 4.1′ with $y(\lambda) = y_\lambda$ being the curve of relative equilibria. Hence the theorem follows from Theorem 4.1′ and part (2) of the proof of Theorem 5.1.

Suppose that Δ is finite and cyclic. Let $Q \subset S(\delta_0)$ be the set of elements which does not generate a Cartan subgroup. We show that Q is a countable union of submanifolds of $S(\delta_0)$, each of codimension greater than or equal to 1. Let K be a Cartan subgroup containing δ_0 and generated by an element of $S(\delta_0)$. The existence of K follows from Proposition 5.4. Recall that K is isomorphic to $K_0 \times Z_l$ with $K_0 \times \{1\}$ corresponding to $\delta_0 K_0$ (we identify Z_l with $\{0, 1, 2, \cdots, l-1\}$). In K_0 we define the sets P^m (see the proof of Proposition 4.3). It follows that $\delta_0 K_0 \cap Q$ is the union of the sets $P^m \times \{1\}$. Let E^m be the union of all conjugacy classes of $P^m \times \{1\}$ by the elements of $(N(\Sigma)/\Sigma)_0$. Proposition 5.5 implies that Q is the union of the sets E^m. Consider the action of $(N(\Sigma)/\Sigma)_0$ on $N(\Sigma)/\Sigma$ defined as conjugation by a group element. As in the proof of Theorem 4.1′ we can partition E^m into manifolds $E^m(\Delta)$, consisting of all elements of E^m whose isotropy with respect to this action is conjugate to Δ.

The remaining part of the proof is analogous to the proof of Theorem 4.1′. The main objective is to show that there exists a residual subset $\mathcal{B} \subset C_T^\infty(R^n \times R, R^n)$ such that if $H \in \mathcal{B}$ then for all choices of m and Δ the curve $\Theta(H)$ is transverse to $E^m(\Delta)$ at every $\lambda \in (0, \lambda_0)$. This is done by showing that for some fixed m and Δ there exists a residual set $\mathcal{B}^m(\Delta)$, whose elements are transverse to $E^m(\Delta)$, and then taking the intersection of the sets $\mathcal{B}^m(\Delta)$.

We now fix m and Δ and regard Θ as a mapping from $C_T^\infty(R^n \times R, R^n)$ to $C^\infty(I, S(\delta_0))$. To show existence of $\mathcal{B}^m(\Delta)$ we need to prove the following properties of Θ:

(1) If the interval I is such that $\bar{I} \subset (0, \lambda_0)$, then Θ is continuous in the Whitney C^∞ topology.

(2) For each $\mu \in (0, \lambda_0)$ there exists an interval I such that $\mu \in I$ and $I \subset (0, \lambda_0)$ and a residual set $\mathcal{A} \subset C_T^\infty(R^n \times R, R^n)$ such that if $H \in \mathcal{A}$ then $\Theta(H)$ is transverse to $E^m(\Delta)$ at all $\lambda \in I$.

Property (1) follows from standard theorems on smooth dependence of solutions of ordinary differential equations on parameters (also see the proof of Theorem 2.2).

We now indicate how to prove property (2). We choose an interval I_0 such that $\mu \in I_0$ and $\bar{I}_0 \subset (0, \lambda_0)$. Given some interval $I \subset I_0$ we define $\mathcal{A}_0$ as the set of elements of $C^\infty(I_0, S(\delta_0))$ which are transverse to $E^m(\Delta)$ at all $\lambda \in I$. Standard transversality theory implies that $\mathcal{A}_0$ is residual in the Whitney C^∞ topology on $C^\infty(I_0, S(\delta_0))$. Let $\mathcal{A} = \Theta^{-1}(\mathcal{A}_0)$. The property (1) implies that $\mathcal{A}$ is an intersection of open sets. If I is small enough then $\mathcal{A}$ is dense in the Whitney C^∞ topology on $C_T^\infty(R^n \times R, R^n)$. To see this, suppose that $\gamma \in C^\infty(I_0, S(\delta_0))$ is a small perturbation of the curve $\Theta(F)$. Then there exists a small perturbation H of the family F such that $\Theta(F+H) = \gamma$. The proof of the existence of H is a straightforward generalization of Lemma 5.6.

The methods used in the proof of Theorem 5.2 can be easily generalized to prove the following proposition.

PROPOSITION 5.7. Suppose that F satisfies (H1) and (H2). Then there exists a residual set $\mathcal{B} \subset C_T^\infty(R^n \times R, R^n)$ such that if $H \in \mathcal{B}$ then the dimension of the trajectories of $F+H$ on the sets Y is greater than or equal to rank $(N(\Sigma_Y)/\Sigma_Y)$.

6. Bifurcations of relative equilibria with $O(2)$ symmetry. In this section we discuss bifurcation problems with symmetry group $O(2)$. Let $\dot{+}$ denote semidirect product. Recall that $O(2) = SO(2) \dot{+} Z_2(\kappa)$, where $Z_2(\kappa) = \{1, \kappa\}$, κ is an orientation reversing element of $O(2)$, and $SO(2)$ is the subgroup of $O(2)$ consisting of orientation preserving rotations. We assume that $O(2)$ acts on R^n and that $F: R^n \times R \to R^n$ is a smooth and equivariant family of vector fields. We restrict our attention to bifurcations of group orbits of equilibria whose isotropy subgroups are either $Z_2(\kappa)$ or D_k, $k \geqq 2$. Here D_k denotes the group of symmetries of a regular k-gon. The groups $Z_2(\kappa)$ or D_k occur as maximal isotropy subgroups for the various irreducible representations of $O(2)$; hence the bifurcations we study can occur as secondary bifurcations from an invariant equilibrium.

Let R_θ denote the rotation of the plane by the angle θ. The groups D_k are generated by the reflection κ and the rotation $R_{2\pi/k}$. Let Z_k denote the cyclic group generated by $R_{2\pi/k}$. We have $D_k = Z_k \dot{+} Z_2(\kappa)$. In the sequel we use D_1 to denote $Z_2(\kappa)$ and Z_1 to denote the trivial subgroup **1**.

Fix $x_0 \in R^n$, let $X = O(2)x_0$, and let Σ_{x_0} be the isotropy subgroup of x_0. We assume that $\Sigma_{x_0} = D_k$, $k \geqq 1$, and that X is a relative equilibrium of F. Let G be as defined in § 3; that is, G is the restriction of F_N to the normal space N_{x_0}. Let $g = G(\cdot, 0)$; and let E be the center subspace of $(dg)_{x_0}$. In this section we analyze steady-state and Hopf bifurcations of F near X. More precisely, we consider the following situations:

(a) $(dg)_{x_0}$ has a zero eigenvalue.

(b) $(dg)_{x_0}$ has a purely imaginary eigenvalue $i\omega$.

Let $\mathcal{K}(\Sigma_{x_0}) = \{\sigma \in \Sigma_{x_0}: \sigma v = v \text{ for all } v \in E\}$ be the *kernel* of the action of Σ_{x_0} on E. We assume that the action of Σ_{x_0} on E is nontrivial; that is, $\mathcal{K}(\Sigma_{x_0})$ is properly contained in Σ_{x_0}. We now state the two main results of this section: a steady-state bifurcation theorem and a Hopf bifurcation theorem. We begin with the steady-state bifurcation theorem. Suppose that $(dg)_{x_0}$ has a zero eigenvalue. We make a generic assumption that E is the nullspace of $(dg)_{x_0}$ and that the action of Σ_{x_0} on E is absolutely irreducible. The following theorem describes all the generic types of bifurcating solutions and gives the number of distinct nonconjugate branches.

THEOREM 6.1. *All the generic types of bifurcating solutions of F are listed in Table* 6.1.

We now state the Hopf bifurcation theorem. Suppose that $(dg)_{x_0}$ has a purely imaginary eigenvalue $i\omega$. We make a generic assumption that E is the eigenspace of $i\omega$ and that the action of Σ_{x_0} on E is Γ-simple. The following theorem describes all the generic types of bifurcating solutions and gives the number of distinct nonconjugate branches.

THEOREM 6.2. *All the generic types of bifurcating solutions of F are listed in Table* 6.2.

TABLE 6.1

Kernel of isotropy	Type of solution	Number of branches
Z_k	rotating wave	1
D_m, $k = 2m$	steady state	1
Z_l, l/k, $l < k$	steady state	2

TABLE 6.2

Kernel of isotropy	Type of solution	Number of branches
Z_k	periodic orbit	1
D_m, $k=2m$	periodic orbit	1
Z_l, l/k, $l<k$	periodic orbit	2
	two-torus	1

In the remainder of this section we prove Theorems 6.1 and 6.2. We begin by classifying all the possible kernels of the action of Σ_{x_0} on E. We prove the following lemma.

LEMMA 6.3. *One of the following statements must hold*:

(i) $\mathcal{K}(\Sigma_{x_0}) = Z_l$, $l \leqq k$, l *divides* k, $l \neq k/2$.

(ii) $k = 2m$ *and* $\mathcal{K}(\Sigma_{x_0})$ *is isomorphic to* D_m.

Proof. Note that $\mathcal{K}(\Sigma_{x_0})$ is normal in Σ_{x_0}. Hence finding all the possible groups $\mathcal{K}(\Sigma_{x_0})$ is equivalent to classifying the normal subgroups of Σ_{x_0}.

We consider two cases: $\mathcal{K}(\Sigma_{x_0}) \subset SO(2)$ and $\mathcal{K}(\Sigma_{x_0}) \not\subset SO(2)$. Suppose that $\mathcal{K}(\Sigma_{x_0}) \subset SO(2)$. Then $\mathcal{K}(\Sigma_{x_0})$ is a subgroup of Z_k. Hence $\mathcal{K}(\Sigma_{x_0}) = Z_l$, $l \leqq k$, l divides k.

Suppose now that $\mathcal{K}(\Sigma_{x_0}) \not\subset SO(2)$. Let $\zeta \in \mathcal{K}(\Sigma_{x_0})$, $\zeta \notin SO(2)$. Since $\mathcal{K}(\Sigma_{x_0})$ is normal in Σ_{x_0} we have $R_{4\pi/k}\zeta = R_{2\pi/k}\zeta R_{-2\pi/k} \in \mathcal{K}(\Sigma_{x_0})$. Hence $R_{4\pi/k} \in \mathcal{K}(\Sigma_{x_0})$. If $k = 2m$ then it follows that $\mathcal{K}(\Sigma_{x_0})$ is generated by $R_{4\pi/k}$ and ζ and therefore is isomorphic to D_m. Otherwise $k+1$ is divisible by 2 and $(R_{4\pi/k})^{(k+1)/2} = R_{2\pi/k}$. This implies that $\mathcal{K}(\Sigma_{x_0}) = D_k = \Sigma_{x_0}$, which is a contradiction.

The case $\mathcal{K}(\Sigma_{x_0}) = Z_l$, $l \neq k/2$ cannot occur, since then $\Sigma_{x_0}/\mathcal{K}(\Sigma_{x_0})$ would be isomorphic to D_2 and every irreducible action of D_2 has a nontrivial kernel.

Proof of Theorem 6.1. We begin by describing the bifurcation problem for the family G. We assume that the center manifold reduction has been carried out; that is, G is a family of Σ_{x_0}-equivariant vector fields on E. Recall that $\mathcal{K}(\Sigma_{x_0})$ is normal in Σ_{x_0}. Let $\tau : \Sigma_{x_0} \to \Sigma_{x_0}/\mathcal{K}(\Sigma_{x_0})$ be the natural projection. We define the action of $\tau(\Sigma_{x_0})$ on E by $\tau(\sigma)v = \sigma v$, $v \in E$, $\sigma \in \Sigma_{x_0}$. Note that this action is well defined, since $\ker \tau = \mathcal{K}(\Sigma_{x_0})$. It follows that G is Σ_{x_0}-equivariant if and only if it is $\tau(\Sigma_{x_0})$-equivariant. We replace the action of Σ_{x_0} by the action of $\tau(\Sigma_{x_0})$.

As indicated in Table 6.1 we divide the analysis into three cases:

(1) $\mathcal{K}(\Sigma_{x_0}) = Z_l$, $l < k$, l divides k.

(2) $\mathcal{K}(\Sigma_{x_0}) = Z_k$.

(3) $k = 2m$ and $\mathcal{K}(\Sigma_{x_0})$ is isomorphic to D_m.

Case (1). Let $m = k/l$. Since $l < k$ it follows that $m \geqq 2$. Clearly, $\tau(\Sigma_{x_0})$ is isomorphic to D_m. Since the action of $\tau(\Sigma_{x_0})$ is faithful it follows that $\dim E > 1$. This implies that $\dim E = 2$, since all the irreducible representations of D_m are one- or two-dimensional. Also $m \geqq 3$, since any irreducible representation of D_2 has a nontrivial kernel. It follows that the action of $\tau(\Sigma_{x_0})$ is isomorphic to the standard action of D_m on C. According to Table XIII, 5.2 in Golubitsky, Stewart, and Schaeffer [1988] a generic family G has two branches of steady-state solutions $y_1(\lambda)$ and $y_2(\lambda)$. Let $Y_1 = O(2)y_1$ and $Y_2 = O(2)y_2$. We now show that the sets Y_1 and Y_2 consist of equilibria of F. The results of Golubitsky, Stewart, and Schaeffer also imply that the isotropy subgroups of y_1 and y_2 with respect to the action of D_m are two-element groups, each generated by an element not contained in Z_m. Let Σ_{y_1} and Σ_{y_2} denote the isotropy subgroups of y_1 and y_2 in Σ_{x_0}. It follows that $\Sigma_{y_1} \not\subset SO(2)$ and $\Sigma_{y_2} \not\subset SO(2)$. Hence the

normalizers of Σ_{y_1} and Σ_{y_2} are discrete. Theorem 4.1 implies that the group orbits Y_1 and Y_2 consist of equilibria of F.

Case (2). Note that $\tau(\Sigma_{x_0})$ is isomorphic to Z_2. Hence dim $E = 1$. Since the action of Z_2 is nontrivial it follows that Z_2 acts on E as minus identity. Hence generically G undergoes a pitchfork bifurcation; that is, G has a unique (up to conjugacy) branch of equilibria $y(\lambda)$ with trivial isotropy. For more information on Z_2-equivariant bifurcation, see Golubitsky and Schaeffer [1985, Chap. XVI].

Let $Y = O(2)y$. We show that generically the trajectories of F on Y are rotating waves. Let Σ_y be the isotropy subgroup of y in Σ_{x_0}. It follows that $\Sigma_y = \mathcal{H}(\Sigma_{x_0}) = Z_k$. We conclude that $N(\Sigma_y) = O(2)$. By Theorem 4.1 generically the trajectories of $F(\cdot, \lambda)$ on $Y(\lambda)$ are given by drift along circles.

Case (3). As in Case (2) $\tau(\Sigma_{x_0})$ is isomorphic to Z_2. Hence G has a branch of equilibria $y(\lambda)$, whose isotropy subgroup in Z_2 is trivial. Let Σ_y be the isotropy subgroup of y in Σ_{x_0}. Since $\Sigma_y = \mathcal{H}(\Sigma_{x_0})$ and $\mathcal{H}(\Sigma_{x_0})$ is isomorphic to D_m, it follows that $\Sigma_y \not\subset SO(2)$ and that $N(\Sigma_y)$ is discrete. By Theorem 4.1 the orbit $Y = O(2)y$ consists of equilibria of F.

Before proving Theorem 6.2 we give some background on Hopf bifurcation from an invariant equilibrium. The results we review will be used in the analysis of bifurcations of the family G. Let $\Gamma \subset O(n)$ be a Lie group acting on R^n and suppose that this action is nontrivial. Let $H: R^n \times R \to R^n$ be a family of smooth, Γ-equivariant vector fields. Let $h = H(\cdot, 0)$ and suppose that $(dh)_0$ has a purely imaginary eigenvalue ωi. Suppose that the center manifold reduction has been carried out; that is, R^n is the real part of the sum of the eigenspaces of $\pm\omega i$. We make a generic assumption that the action of Γ on R^n is Γ-simple. Then the group $\{\exp(Lt): t \in R\}$ is isomorphic to S^1. We define the action of S^1 on R^n as

$$(\gamma, \theta)x = \gamma \exp(L\theta)x \quad \text{where } (\gamma, \theta) \in \Gamma \times S^1 \text{ and } x \in R^n.$$

The following theorem is the equivariant Hopf bifurcation theorem (see Golubitsky and Stewart [1985, Thm. 5.1] or Golubitsky, Stewart, and Schaeffer [1988, Thm. XVI, 4.1]):

THEOREM 6.4. *Suppose that* Δ *is a maximal isotropy subgroup of* $\Gamma \times S^1$ *and* dim Fix $(\Delta) = 2$. *Then* H *has a branch of small amplitude periodic solutions* $x_\lambda(t)$ *such that* $\sigma x_\lambda(t + \theta) = x_\lambda(t)$ *for every pair* $(\sigma, \theta) \in \Delta$.

Suppose that $x_\lambda(t)$ is a branch of periodic solutions described in the statement of Theorem 6.4. Let $X(\lambda) = \{x_\lambda(t): t \in R\}$. Recall the definition of the group of symmetries of the set X, denoted by Σ_X, as the set of all $\sigma \in \Gamma$ such that $\sigma x = x$ for all $x \in R^n$. Clearly, Σ_X is obtained by projecting Δ onto the first component of $\Gamma \times S^1$; that is,

$$\Sigma_X = \{\sigma \in \Gamma: (\sigma, \theta) \in \Delta \text{ for some } \theta \in S^1\}.$$

We refer to the group Δ as the group of spatial-temporal symmetries of the periodic orbit X.

We will now describe generic Hopf bifurcations in two cases: $\Gamma = Z_2$ and $\Gamma = D_k$, $k \geqq 3$. Assume that $\Gamma = Z_2$. We have the following proposition.

PROPOSITION 6.5. *Generically, the family* H *has a branch of periodic orbits* $Y(\lambda)$ *with* $\Sigma_{Y(\lambda)} = Z_2$.

Proof. The irreducible representations of Z_2 are absolutely irreducible and one-dimensional. Hence a Γ-simple representation of Z_2 will be two-dimensional and will have the form $R \oplus R$. The action on each of the copies of R will be given as reflection with respect to the origin. Let ζ be the nontrivial element in Z_2. Then, for $(x, y) \in R^2$

$\zeta(x, y) = (-x, -y)$. Also

$$L = \begin{pmatrix} 0 & -\omega \\ \omega & 0 \end{pmatrix}$$

and $\exp(Lt)$ is a rotation by angle t. It is easy to see that $Z_2(\zeta, \pi) = \{(0, 0), (\zeta, \pi)\}$ is a maximal isotropy subgroup of $Z_2 \times S^1$. It follows from Theorem 6.4 that H has a branch of periodic orbits $Y(\lambda)$ with $\Sigma_{Y(\lambda)} = Z_2$. Using normal form theory we can show that generically this branch is unique.

We now discuss the case where $\Gamma = D_k$, $k \geqq 3$. We assume that the action of D_k on R^n is faithful. Let $\xi = 2\pi/k$. We define the following subgroups of $D_k \times S^1$:

$$\tilde{Z}_k = \left\{ \left(\frac{2\pi m}{k}, \frac{2\pi m}{k} \right) : m = 0, 1, \cdots, k \right\},$$

$$Z_2(\kappa) = \{(0, 0), (\kappa, 0)\},$$

$$Z_2(\kappa, \pi) = \{(0, 0), (\kappa, \pi)\},$$

$$Z_2(\kappa, \xi) = \{(0, 0), (\kappa, \xi)\},$$

and when k is even

$$Z_2^c = \{(0, 0), (\pi, \pi)\}.$$

We have the following theorem.

THEOREM 6.6. *Generically, the family H has three branches of periodic orbits*: $Y_1(\lambda)$, $Y_2(\lambda)$, *and* $Y_3(\lambda)$. *The groups of spatial-temporal symmetries of* $Y_1(\lambda)$, $Y_2(\lambda)$, *and* $Y_3(\lambda)$ *are given respectively by*:

(a) $\tilde{Z}_k$, $Z_2(\kappa)$, *and* $Z_2(\kappa, \pi)$ *if k is odd.*

(b) $\tilde{Z}_k$, $Z_2(\kappa) \oplus Z_2^c$, *and* $Z_2(\kappa, \pi) \oplus Z_2$ *if* $k \equiv 2 \pmod 4$.

(c) $\tilde{Z}_k$, $Z_2(\kappa) \oplus Z_2^c$, *and* $Z_2(\kappa, \xi) \oplus Z_2$ *if* $k \equiv 0 \pmod 4$.

Theorem 6.6 is a consequence of Theorem XVIII, 3.1 in Golubitsky, Stewart, and Schaeffer [1988].

We now present the proof of Theorem 6.2.

Proof of Theorem 6.2. We begin by describing the bifurcation problem for the family G. We assume that the center manifold reduction has been carried out; that is, G is a family of Σ_{x_0}-equivariant vector fields on E. Recall that τ is the natural projection from Σ_{x_0} onto $\Sigma_{x_0}/\mathcal{K}(\Sigma_{x_0})$. As in the proof of Theorem 6.1 we replace the action of Σ_{x_0} by the action of $\tau(\Sigma_{x_0})$. We consider three cases:

(1) $\mathcal{K}(\Sigma_{x_0}) = Z_l$, $l < k$, l divides k.

(2) $\mathcal{K}(\Sigma_{x_0}) = Z_k$.

(3) $k = 2m$ and $\mathcal{K}(\Sigma_{x_0})$ is isomorphic to D_m.

Case (1). Let $m = l/k$. We have $\tau(\Sigma_{x_0}) = D_m$. Since the action of $\tau(\Sigma_{x_0})$ on E is faithful it follows that $m \geqq 3$. The action of $\tau(\Sigma_{x_0})$ on E is Γ-simple, so we are in position to apply Theorem 6.6. Hence G has three branches of solutions $Y_1(\lambda)$, $Y_2(\lambda)$, and $Y_3(\lambda)$ whose groups of spatial-temporal symmetries in D_m are as indicated in Theorem 6.6. Let Σ_{Y_1}, Σ_{Y_2}, and Σ_{Y_3} be the groups of symmetries of these trajectories inside of Σ_{x_0}. Observe that $\Sigma_{Y_1} = Z_k$ and the groups Σ_{Y_2} and Σ_{Y_3} are not contained in $SO(2)$. It follows that $N(\Sigma_{Y_1}) = O(2)$ and the normalizers of the groups Σ_{Y_2} and Σ_{Y_3} are discrete. Let $Z_1(\lambda) = O(2) Y_1(\lambda)$, $Z_2(\lambda) = O(2) Y_2(\lambda)$, and $Z_3(\lambda) = O(2) Y_3(\lambda)$. Theorem 5.2 implies that generically the trajectories of F on Z_1 are dense in two-dimensional tori and the trajectories of F on Z_2 and Z_3 are periodic orbits.

Case (2). We have $\tau(\Sigma_{x_0}) = Z_2$. By Proposition 6.5 G has a unique branch of periodic orbits $Y(\lambda)$. Let ζ be the nontrivial element in Z_2. It follows from the proof of Proposition 6.5 that (ζ, π) is a spatial-temporal symmetry of Y. Let Σ_Y denote the group of symmetries of Y inside of Σ_{x_0}. We have $\tau(\kappa) = \zeta$, so $\Sigma_Y \not\subset SO(2)$ and $N(\Sigma_Y)$ is discrete. Let $Z = O(2)Y$. It follows that the flow of F on Z consists of periodic orbits.

Case (3). We have $\tau(\Sigma_{x_0}) = Z_2$. By Proposition 6.5 G has a unique branch of periodic orbits $Y(\lambda)$. Since $\mathcal{K}(\Sigma_{x_0}) \not\subset SO(2)$ it follows that $\Sigma_Y \not\subset SO(2)$ and $N(\Sigma_Y)$ is discrete. Let $Z = O(2)Y$. It follows that the flow of F on Z consists of periodic orbits.

7. The Kuramoto–Sivashinsky equation. The Kuramoto-Sivashinsky equation is used to model several physical and chemical phenomena, for example, flame propagation and some aspects of the dynamics of the Belousov-Zhabotynski reaction. The following is the Kuramoto-Shivashinsky equation in one space variable:

$$u_t + 4u_{xxxx} + \alpha(u_{xx} + \tfrac{1}{2}u_x^2) = 0. \tag{7.1}$$

In this section we study a bifurcation problem derived from (7.1). Equation (7.1) is equivariant with respect to translations and reflections in the space variable. An approach often used in such situations is to impose periodic boundary conditions with period $L > 0$. Then L becomes an additional parameter in the problem. The space variable x can be rescaled so that the boundary conditions become 2π periodic. As a result we obtain the following boundary value problem:

$$v_t + 4v_{xxxx} + \alpha(v_{xx} + \tfrac{1}{2}v_x^2) = 0, \qquad v(x + 2\pi, t) = v(x, t) \tag{7.2}$$

where the period L has been absorbed into the parameter α. The boundary value problem (7.2) is $O(2)$-equivariant. Hence the theory developed in § 6 will apply to bifurcations of relative equilibria of (7.2).

An interesting aspect of the bifurcation analysis of the Kuramoto-Sivashinsky equation is that we can easily find primary branches of solutions with isotropy D_k, for all $k \geqq 2$. This is a consequence of the following observation. Suppose that u is a steady-state solution of (7.2). If we extend u by periodicity to the interval $[0, 2k\pi]$ and rescale the space variable by k, then the so-obtained function is an equilibrium solution of (7.2) for a different value of the parameter α. The new equilibrium is called a *replicated solution.* Note that this solution is $2\pi/k$ periodic, which implies that its isotropy subgroup contains Z_k. It is easy to see that $u = 0$ is an equilibrium of (7.2). This equilibrium is stable for α near zero. As α is increased the solution $u = 0$ loses stability and bifurcates to a branch of solutions with isotropy group $Z_2(\kappa)$. Hence for each $k \geqq 2$ there exists a branch of replicated equilibria with symmetry D_k.

We might expect that the secondary branches of solutions bifurcating along the replicated branches would be replications of the secondary branches bifurcating from the primary branch. According to the analysis of § 6, however, secondary bifurcations from the replicated branches can be different from secondary bifurcations from the branch with symmetry $Z_2(\kappa)$. In particular, we expect the branch with isotropy $Z_2(\kappa)$ to bifurcate to a rotating wave, and the branches with isotropy D_k to bifurcate to group orbits of equilibria. Kevrekedis, Nicolaenco, and Scovel [1988] carried out a computer-assisted study of secondary and tertiary bifurcations from the branches with isotropy groups $Z_2(\kappa)$, D_2, and D_3. Their results fit the predictions of Theorems 6.1 and 6.2; in particular, the first bifurcation along the branch of the equilibria with isotropy group $Z_2(\kappa)$ is to a rotating wave, and the first bifurcations along the branches of equilibria with isotropy groups D_2 and D_3 are to orbits of equilibria. In this section we discuss

the results of Kevrekedis, Nicolaenco, and Scovel and compare them with the predictions of Theorems 6.1 and 6.2.

The numerical results of Kevrekedis, Nicolaenco, and Scovel also indicate existence of quasi-periodic solutions and dynamics related to homoclinic and heteroclinic connections. None of these arise as a result of the bifurcations discussed in § 6. Armbruster, Guckenheimer, and Holmes [1987] analyzed the $O(2)$ equivariant problem of interaction of two steady-state modes, one with isotropy group $Z_2(\kappa)$ and the other with isotropy group D_2. The dynamics they found was much like the dynamics found by Kevrekedis, Nicolaenco, and Scovel near the $Z_2(\kappa)$ and D_2 branches.

Let us now give a more detailed description of the bifurcation problem derived from the Kuramoto–Sivashinsky equation. We start by modifying the coordinates in (7.2) so that the solutions are bounded (see Kevrekedis et al. [1988]):

$$m(t) = \int_0^{2\pi} v(x, t)\, dx.$$

We use (7.2) and the fact that the integrals $\int_0^{2\pi} v_{xx}\, dx$ and $\int_0^{2\pi} v_{xxxx}\, dx$ vanish to show that

$$\dot{m}(t) = -\frac{\alpha}{4\pi} \int_0^{2\pi} v_x^2\, dx.$$

We now modify the coordinates by letting $u(x, t) = v(x, t) - m(t)$. We obtain

$$\begin{aligned} &u_t + 4u_{xxxx} + \alpha(u_{xx} + \tfrac{1}{2}u_x^2) + m(t) = 0, \\ &u(x, t) = u(x + 2\pi, t). \end{aligned} \tag{7.3}$$

Let us now describe the symmetries of (7.2) and (7.3). Let X be a space of four times differentiable functions $u(x, t)$, 2π periodic in the space variable x. The $O(2)$ action on X is generated by

$$\begin{aligned} &\theta u(x, t) = u(x + \theta, t), \qquad \theta \in SO(2), \\ &\kappa u(x, t) = u(-x, t). \end{aligned}$$

It is easy to see that (7.2) and (7.3) are equivariant with respect to this action.

Let us now explain in more detail how we obtain the replicated steady-state solutions. The ideas we present can be found in Kevrekedis, Nicolaenco, and Scovel. Consider the steady-state problem corresponding to (7.3):

$$4u_{xxxx} + \alpha\left(u_{xx} + \frac{1}{2}u_x^2\right) + \int_0^{2\pi} u_x^2\, dx = 0. \tag{7.4}$$

We assert the following. Suppose $u(x)$ is a steady-state solution of (7.3) with $\alpha = \alpha_0$ and let k be a positive integer. Then $w(x) = u(kx)$ is a solution of (7.3) with $\alpha = 4k^2\alpha_0$. To prove the assertion we apply the left-hand side of (7.3) to w and use the fact that u is a solution.

Note that $u = 0$ is a trivial solution of (7.2). To determine the stability of zero we write (7.2) as

$$u_t = F(u)$$

where

$$F(u) = -4u_{xxxx} + \alpha(u_{xx} + \tfrac{1}{2}u_x^2) + \dot{m}(t).$$

Then

$$dF|_{u=0}h = -4h_{xxxx} + \alpha h_{xx}.$$

It is easy to see that the functions $e^{2\pi ikx}$, where k is an integer, form a complete set of eigenvectors of $(dF)_{u=0}$ and the corresponding eigenvalues are $(2\pi)^2k^2(4k^2-\alpha)$. The first instability occurs at $k=1$ and $\alpha=4$. As α crosses 4, a branch of equilibria with isotropy group $Z_2(\kappa)$ bifurcates from the trivial solution. We will refer to it as the unimodal branch. This branch is replicated for $\alpha=4k^2$. These replicated branches will be referred to as the k-modal branches. Kevrekedis, Nicolaenco, and Scovel describe some secondary and tertiary bifurcations discovered in their numerical studies of the Kuramoto-Sivashinsky equation. If we believe that those bifurcations are generic in the sense discussed in § 6, then each one of them must match one of the cases described in Theorems 6.1 and 6.2. In what follows we summarize the findings of Kevrekedis, Nicolaenco, and Scovel and relate them to the results of Theorems 6.1 and 6.2. The bifurcation diagram based on the results of Kevrekedis, Nicolaenco, and Scovel is given in Fig. 7.1. The solid lines represent branches of asymptotically stable solutions and the dotted ones represent branches of unstable solutions.

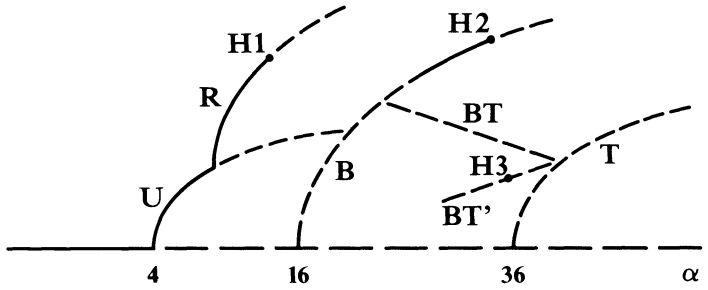


FIG. 7.1. *Secondary and tertiary bifurcation of the Kuramoto-Shivashinsky equation.*

Let U be the branch of equilibria with isotropy $Z_2(\kappa)$, B the branch of equilibria with isotropy D_2, and T the branch of equilibria with isotropy D_3. We discuss the secondary bifurcations found by Kevrekedis, Nicolaenco, and Scovel along each of these branches. We first consider steady-state bifurcations.

(1) *Steady-state bifurcations from the branch U.* At $\alpha=13.005$ the computations of Kevrekedis, Nicolaenco, and Scovel reveal that a real eigenvalue passes through zero. In this case Theorem 6.1 predicts a bifurcation to a rotating wave. The numerical experiments confirm the existence of a rotating wave. Moreover, Kevrekedis, Nicolaenco, and Scovel give an analytical proof of the existence of the rotating wave, based on the ideas of Iooss [1986].

(2) *Steady-state bifurcations from the branch B.*

(a) The first bifurcation on the branch B occurs at $\alpha=16.1399$. Theorem 6.1 predicts a bifurcation to a unique branch of orbits of equilibria with isotropy group isomorphic to Z_2 (provided that the kernel of the action of D_2 on the nullspace is not contained in $SO(2)$). This is in agreement with the computations, which show that the branch U merges with the branch B.

(b) An analogous bifurcation is observed for $\alpha=22.559$. The isotropy group bifurcating branch is $Z_2(\kappa)$. We label this branch BT since it later joins with the trimodal branch.

(3) *Steady-state bifurcations from the branch T.* At $\alpha=36.235$ two real eigenvalues of the branch T pass through zero. According to Theorem 6.1 there are two branches of equilibria bifurcating of the branch T. The isotropy of these equilibria is $Z_2(\kappa)$. The

numerical results are in complete agreement with this prediction. One of the bifurcating branches is the branch *BT*. Kevrekedis, Nicolaenco, and Scovel refer to the other branch as a continuation of *BT* and also label it *BT*. For this reason we label this branch as *BT'*.

(4) *Steady-state bifurcations from the branch BT.* Kevrekedis, Nicolaenco, and Scovel also find a bifurcation point related to a zero eigenvalue on the branch *BT*. They conjecture that the corresponding bifurcation is to a rotating wave. Theorem 6.1 also predicts a bifurcation to a rotating wave and hence supports the conjecture.

Kevrekedis, Nicolaenco, and Scovel discuss three Hopf bifurcation points, marked in Fig. 7.1 as H1, H2, and H3. The following are the predictions of the nature of these bifurcations derived from Theorem 6.2.

The Hopf bifurcation point 1. Point H1 corresponds to a Hopf bifurcation from the branch of rotating waves *R*. The group of symmetries of the branch of rotating waves is $SO(2)$. Theorem 6.2 implies that generically the bifurcating trajectories are dense in two-tori. This agrees with the predictions of Kevrekedis, Nicolaenco, and Scovel [1988, closing remarks of § 5a], who conclude from the structure of the dynamics that a doubly periodic solution is likely to exist.

The Hopf bifurcation point 2. Point H2 corresponds to a Hopf bifurcation occurring along the branch *B*. The isotropy group of the equilibria on the branch *B* is D_2. Theorem 6.2 implies that this Hopf bifurcation leads to a periodic flow. The numerical results of Kevrekedis, Nicolaenco, and Scovel indicate that the bifurcating solutions are periodic.

The Hopf bifurcation point 3. Point H3 corresponds to Hopf bifurcation occurring along the branch *BT*. Let Σ be the isotropy group of the equilibria on that branch. We have previously remarked that $\Sigma = Z_2(\kappa)$ or $\Sigma = Z_2(\kappa, \pi)$. It follows from Theorem 6.2 that the bifurcating solutions must be periodic orbits. Kevrekedis, Nicolaenco, and Scovel do not comment on the dynamics related to this bifurcation.

8. The Bénard problem. In this section we analyze secondary steady-state bifurcations of a dynamical system equivariant with respect to the group $\Gamma = D_6 \dot{+} \mathbb{T}^2$, where $\mathbb{T}^2$ is a two-dimensional torus and D_6 is the group of symmetries of a regular hexagon. A bifurcation problem with this symmetry arises in the analysis of the mathematical model of convection between two infinite planes. This problem is called the planar Bénard problem. We now briefly describe the symmetries of the model and the derivation of the bifurcation problem with symmetry group Γ. Detailed information on this topic and the analysis of primary bifurcations can be found in Buzano and Golubitsky [1983] or in Golubitsky, Stewart, and Schaeffer [1988, Case Study 4].

Let x, y be the coordinates in a horizontal plane and z the coordinate in the vertical direction. The model of convection is equivariant with respect to translations, reflections, and rotations in the xy plane. The group generated by these transformations is called the group of Euclidian motions in the plane and is denoted by E_2. Let w be in R^2 and let T_w denote translation by w. The group of translations in the plane is isomorphic to R^2 with the isomorphism defined by the assignment $w \leftrightarrow T_w$. The group E_2 can be thought of as the semidirect product $O(2) \dot{+} R^2$ with multiplication defined by

$$(8.1) \qquad (\sigma_1, T_{w_1})(\sigma_2, T_{w_2}) = (\sigma_1\sigma_2, T_{w_1+\sigma_1 w_2}), \qquad \sigma_1, \sigma_2 \in O(2), \quad w_1, w_2 \in R^2.$$

Let $e \in R^2$ be an arbitrary vector and let f be obtained by rotating e by $\pi/3$. The hexagonal lattice H_6 is given as

$$H_6 = \{ne + mf: \text{ for all pairs of integers } n \text{ and } m\}.$$

Note that H_6 is a subgroup of R^2. The solutions of the convection problem have the form $u(t, x, y, z)$. We restrict our attention to those that are periodic in both directions of the lattice. Let

$$\mathscr{X} = \{u(t, x, y, z): u(t, (x, y, z) + (e, 0)) = u(t, (x, y, z) + (f, 0)) = u(t, x, y, z)\}.$$

Clearly, the only elements of $O(2)$ that leave $\mathscr{X}$ invariant are the elements of D_6. Also the action of H_6 on $\mathscr{X}$ is trivial. Hence the group of symmetries of the convection problem restricted to $\mathscr{X}$ is given by Γ with $\mathbb{T}^2 = R^2/H_6$.

Multiplication in Γ is induced by multiplication in E_2. Let D_6 act on R^2 by the standard action. Let 1 denote the identity in D_6, zero the identity in $\mathbb{T}^2$, and e the identity in Γ. For $p \in R^2$ let p' denote the image of p under the natural projection $R^2 \to \mathbb{T}^2$. For $\sigma \in D_6$ we define $\sigma \cdot p' = (\sigma p)'$. Multiplication in Γ is given as follows:

$$(\sigma_1, p_1')(\sigma_2, p_2') = (\sigma_1\sigma_2, p_1' + \sigma_1 \cdot p_2'). \tag{8.2}$$

Here $\sigma_1\sigma_2$ is the product of σ_1 and σ_2 in D_6 and $p_1' + \sigma_1 \cdot p_2'$ is the sum of vectors in $\mathbb{T}^2$.

The Bénard problem has an invariant equilibrium (the pure conduction state). There exists a region in the parameter space where this equilibrium is stable. The known primary bifurcations are to two types of equilibria with maximal isotropy subgroups. These subgroups are D_6 and $D_2 \dot{+} S^1$ $(= Z_2 \oplus O(2))$. The equilibria with isotropy D_6 are called hexagons, and the equilibria with isotropy $Z_2 \oplus O(2)$ are called rolls. In what follows we describe the kinds of steady-state bifurcations each one of these solutions can undergo.

From now on we consider an abstract Γ-equivariant bifurcation problem. We assume that $F: R^n \to R^n \times R$ (for some n) is a smooth Γ-equivariant family of vector fields and that F has a branch of equilibria with isotropy group D_6 (which we refer to as hexagons) and a branch of solutions with isotropy group $Z_2 \oplus O(2)$ (which we refer to as rolls). We analyze the generic bifurcations of these solutions.

(A) Bifurcations of hexagons. Suppose $X = \Gamma x_0$ is a group orbit of hexagons. Let G be the restriction of F_N to the normal space $x_0 + N_{x_0}$ and let $g = G(\cdot, 0)$. We assume that $(dg)_{x_0}$ has a zero eigenvalue. Let E be the center subspace of $(dg)_{x_0}$. We make a generic assumption that E is the nullspace of $(dg)_{x_0}$ and that the action of D_6 on E is absolutely irreducible. Our bifurcation analysis will depend on the action of D_6 on E. Let $\mathscr{K}(D_6)$ be the kernel of the action of D_6 on E. We assume that the bifurcation is symmetry breaking, that is, $\mathscr{K}(D_6)$ is a proper subgroup of D_6. According to Lemma 6.3, either $\mathscr{K}(D_6) = Z_m$, $m = 1, 2, 6$, or $\mathscr{K}(D_6)$ is isomorphic to D_3. The following proposition gives a classification of generic bifurcations of hexagons.

PROPOSITION 8.1. *All generic types of bifurcating solutions of F are listed in Table* 8.1.

We now state and prove a lemma necessary to prove Proposition 8.1. Suppose that Σ is a subgroup of D_6 and let $N(\Sigma)$ denote the normalizer of Σ in Γ. Let Fix (Σ)

TABLE 8.1

Kernel of isotropy	Type of solution	Number of half branches
Z_6 or D_3	steady state	1
Z_2	steady state	2
trivial	periodic orbit	2

denote the fixed-point space of Σ taken with respect to the standard action of D_6 on R^2. We have the following lemma.

LEMMA 8.2. $\dim N(\Sigma) = \dim \mathrm{Fix}(\Sigma)$.

Proof. Let Γ_0 denote the connected component of the identity in Γ. Since Γ is a compact group it suffices to show that the normalizer of Σ in Γ_0 has the same dimension as $\mathrm{Fix}(\Sigma)$. The group Γ_0 consists of elements of form $(1, p')$, $p' \in \mathbb{T}^2$. Suppose that $(\sigma, 0)$ is in Σ. The element $(1, p')$ is in $N(\Sigma)$ if $(1, -p')(\sigma, 0)(1, p')$ is in Σ. We have

$$(1, -p')(\sigma, 0)(1, p') = (\sigma, \sigma \cdot p' - p'). \tag{8.3}$$

If (8.1) holds then we must have $\sigma \cdot p' - p' = 0$, which is equivalent to $(\sigma p - p)' = 0$. The proof now follows.

Proof of Proposition 8.1. Let $\tau: D_6 \to D_6/\mathcal{K}(D_6)$ be the natural projection. In the analysis of bifurcations of the family G we replace the action of D_6 by the action of $\tau(D_6)$. As indicated in Table 8.1 we divide the analysis into three cases:

(1) $\mathcal{K}(D_6) = Z_6$ or $\mathcal{K}(D_6)$ is isomorphic to D_3.

(2) $\mathcal{K}(D_6) = Z_2$.

(3) $\mathcal{K}(D_6)$ is trivial.

Case (1). Observe that $\tau(D_6)$ is isomorphic to Z_2. In this case a generic family G has a unique branch of equilibria $y(\lambda)$, whose isotropy subgroup in D_6 is trivial (see the proof of Theorem 6.1). The isotropy group of $y(\lambda)$ in $\tau(D_6)$ is trivial. Let Σ_y be the isotropy group of $y(\lambda)$ in D_6. It follows that Σ_y is $\mathcal{K}(D_6)$. The group $\mathcal{K}(D_6)$ contains a nontrivial rotation. It follows that the fixed-point space of $\mathcal{K}(D_6)$ with respect to the standard action of D_6 on R^2 is trivial. By Lemma 8.2 $\dim N(\Sigma) = 0$. Let $Y(\lambda) = \Gamma y(\lambda)$. By Theorem 4.1 the set $Y(\lambda)$ consists of equilibria of F.

Case (2). Observe that $\tau(D_6)$ is isomorphic to D_3. It follows from Table XIII, 5.1 in Golubitsky, Stewart, and Schaeffer [1988] that a generic family G has two half branches of equilibria $y_1(\lambda)$, $y_2(\lambda)$. Let Σ_{y_1} and Σ_{y_2} denote the isotropy subgroups of y_1 and y_2 in D_6. Both of these groups must contain $\mathcal{K}(D_6)$. Since $\mathcal{K}(D_6)$ contains the rotation by π it follows that the fixed-point spaces of Σ_{y_1} and Σ_{y_2} are trivial. Lemma 8.2 implies that the normalizers of these groups are discrete. Let $Y_1(\lambda) = \Gamma y_1(\lambda)$, $Y_2(\lambda) = \Gamma y_2(\lambda)$. Theorem 4.1 implies that the sets $Y_1(\lambda)$ and $Y_2(\lambda)$ consist of equilibria of F.

Case (3). According to Table XIII, 5.2 in Golubitsky, Stewart, and Schaeffer [1988], a generic family G has two half branches of equilibria $y_1(\lambda)$, $y_2(\lambda)$, with $\Sigma_{y_1} = \{1, \kappa\}$ and $\Sigma_{y_2} = \{1, \kappa\pi\}$. Here κ denotes the reflection through the x-axis and $\kappa\pi$ denotes the reflection through the y-axis. Clearly, these groups have one-dimensional fixed-point spaces, hence, by Lemma 8.2, their normalizers are one-dimensional. Let $Y_1(\lambda) = \Gamma y_1(\lambda)$, $Y_2(\lambda) = \Gamma y_2(\lambda)$. Theorem 4.1 implies that generically the trajectories of the flow of F on the sets $Y_1(\lambda)$ and $Y_2(\lambda)$ are rotating waves.

(B) Bifurcations of rolls. Suppose $X = \Gamma x_0$ is a group orbit of rolls. Let G be the restriction of F_N to the normal space N_{x_0} and let $g = G(\cdot, 0)$. We assume that $(dg)_{x_0}$ has an eigenvalue on the imaginary axis. Let Σ_r denote the isotropy subgroup of rolls. Let E be the center subspace of $(dg)_{x_0}$. We make a generic assumption that the action of Σ_r on E is irreducible. We will show that this implies that the action of Σ_r on E is absolutely irreducible. This implies that $(dg)_{x_0}$ must have a zero eigenvalue and hence the bifurcation we consider is a steady-state bifurcation.

The group Σ_r is generated by translations along the y-axis, reflection through the y-axis and rotation by π (see Golubitsky, Stewart, and Schaeffer [1988, p. 154]). The projection of the y-axis into the torus $\mathbb{T}^2$ is a circle which we denote by S^1. Let ξ correspond to the element of D_6 which acts on R^2 as rotation by π, and let ζ denote

the element of D_6 which acts as reflection through the y-axis. The element ζ commutes with the elements of S^1 and the element ξ anticommutes with the elements of S^1. It follows that Σ_r is isomorphic to $Z_2 \oplus O(2)$, with S^1 corresponding to $SO(2)$, ξ corresponding to a reflection in $O(2)$, and ζ corresponding to the nontrivial element of Z_2. The irreducible representations of $Z_2 \oplus O(2)$ are given by the irreducible representations of $O(2)$, with ζ acting as identity or as minus identity. Since the irreducible representations of $O(2)$ are absolutely irreducible it follows that $(dg)_{x_0}|E \equiv 0$.

We now divide that analysis into two cases:

(1) S^1 acts trivially on E.

(2) S^1 acts nontrivially on E.

Case (1). In the analysis of bifurcations of G we can replace the action of Σ_r on E by the action of $D_2 = \{1, \xi, \zeta, \xi\zeta\}$. The space E must be one-dimensional and the kernel of the action of D_2 must be one of the groups:

$$Z_2(\zeta) = \{1, \zeta\}, \qquad Z_2(\xi) = \{1, \xi\}, \qquad Z_2(\xi\zeta) = \{1, \xi\zeta\}. \tag{8.4}$$

It follows that in the analysis of bifurcations of G the action of D_2 can be replaced by the action of a group isomorphic to Z_2. Hence a generic family G has a unique branch of equilibria $y(\lambda)$ and the isotropy group of $y(\lambda)$ in D_2 is equal to one of the groups listed in (8.4). Let Σ_y be the isotropy subgroup of $y(\lambda)$ in Σ_r. We have $\Sigma_y = \langle S^1, \zeta \rangle$, or $\Sigma_y = \langle S^1, \xi \rangle$ or $\Sigma_y = \langle S^1, \xi\zeta \rangle$. In other words, Σ_y is generated by the elements of S^1 and ξ, ζ, or $\xi\zeta$. Let $Y(\lambda) = \Gamma y(\lambda)$. We have the following proposition.

PROPOSITION 8.3. *If $\Sigma_y = \langle S^1, \xi\zeta \rangle$ then the trajectories of flow of F on $Y(\lambda)$ are rotating waves. Otherwise, $Y(\lambda)$ consists of equilibria.*

Proof. We prove that if $\Sigma_y = \langle S^1, \xi\zeta \rangle$ then $\dim N(\Sigma_y) = 2$ and otherwise $\dim N(\Sigma_y) = 1$. The proposition will then follow from Theorem 4.1.

By compactness of Γ it suffices to show that the normalizer of Σ_y in Γ_0 has the indicated dimension. Recall that Γ_0 consists of the elements of Γ of the form $(1, p')$, $p' \in \mathbb{T}^2$. Suppose that $(\sigma, 0) \in \Sigma_y$, $\sigma \in D_2$. The identity (8.3) implies that

$$\sigma \cdot p' - p' \in S^1. \tag{8.5}$$

Recall that S^1 is the image of the y-axis under the natural projection $R^2 \to \mathbb{T}^2$. Hence (8.5) implies that $\sigma p - p = (0, q)$, $q \in R$.

Suppose that $p = (p_1, p_2)$. The element $\xi\zeta$ acts on R^2 as reflection through the x-axis; that is, $\xi\zeta(p_1, p_2) = (p_1, -p_2)$. It follows that $\xi\zeta p - p = (0, -2p_2)$. Hence (8.4) holds for all p' in S^1. It follows that if $\Sigma_y = \langle S^1, \xi\zeta \rangle$, then $\dim N(\Sigma_y) = 2$.

The element ξ acts on R^2 as rotation by π; that is, $\xi(p_1, p_2) = (-p_1, -p_2)$. It follows that $\xi p - p = (-2p_1, -2p_2)$. Hence (8.5) holds if $p' \in S^1$. It follows that if $\Sigma_y = \langle S^1, \xi \rangle$, then $\dim N(\Sigma_y) = 1$.

The element ζ acts on R^2 as reflection through the x-axis, that is, $\zeta(p_1, p_2) = (-p_1, p_2)$. It follows that $\zeta p - p = (-2p_1, 0)$. Hence (8.5) holds if $p' \in S^1$. It follows that if $\Sigma_y = \langle S^1, \zeta \rangle$, then $\dim N(\Sigma_y) = 1$.

Case (2). We show that if S^1 acts nontrivially, then generically the flow on the bifurcating relative equilibria is trivial. Namely, we prove the following proposition.

PROPOSITION 8.4. *A generic family F has a unique branch of bifurcating relative equilibria $Y(\lambda)$. The flow on the sets $Y(\lambda)$ is trivial.*

Proof. We first consider bifurcations of the family G. Let $\mathcal{K}(\Sigma_r)$ be the kernel of the action of Σ_r on E. We show that $\mathcal{K}(\Sigma_r)$ must be nontrivial. If ζ acts on E as identity, then $(\zeta, 0) \in \mathcal{K}(\Sigma_r)$, which implies the assertion. Suppose that ζ acts on E as minus identity. Let p_0 be the element of S^1 acting on E as rotation by π. Then $(\zeta, p_0) \in \mathcal{K}(\Sigma_r)$, so $\mathcal{K}(\Sigma_r)$ is nontrivial.

Since the action of S^1 is nontrivial it follows that the action of ξ on E is a reflection. Hence $\mathcal{K}(\Sigma_r)$ is generated by a cyclic subgroup of S^1 and either $(\zeta, 0)$ or (ζ, p_0). It follows that $\Sigma_r/\mathcal{K}(\Sigma_r)$ is isomorphic to $O(2)$. In the analysis of the bifurcations of G we replace the action of Σ_r on E by the action of $\Sigma_r/\mathcal{K}(\Sigma_r)$ on E. By standard results on $O(2)$-equivariant bifurcation G has a unique branch of equilibria $y(\lambda)$ and Σ_y contains a reflection. Since ζ acts as reflection we can assume (by possibly replacing $y(\lambda)$ by a conjugate branch) that $(\zeta, 0) \in \Sigma_y$.

We now prove that the normalizer of Σ_y in Γ_0 is discrete. As we argued earlier, this implies that the normalizer of Σ_y in Γ is discrete. The group Σ_y is generated by $(\zeta, 0)$ and the elements of $\mathcal{K}(\Sigma_r)$. Suppose that $(1, p') \in N(\Sigma_y)$, $p = (p_1, p_2)$. Recall that ξ acts on R^2 as rotation by π; that is, $\xi(p_1, p_2) = (-p_1, -p_2)$. The identity (8.3) implies that $\zeta \cdot p' - p' \in \Sigma_y$. Also $\xi p - p = (-2p_1, -2p_2) = -2p'$. It follows that $2p'$ must be in $\mathcal{K}(\Sigma_r) \cap S^1$, which, by assumption is a discrete group. It follows that $N(\Sigma_y)$ is discrete and by Theorem 4.1 the relative equilibria $Y(\lambda) = \Gamma y(\lambda)$ consist of equilibria of F.

Acknowledgments. The author thanks Marty Golubitsky for very helpful suggestions and comments. The author is also grateful to Mike Field and André Vanderbauwhede for helpful discussions.

REFERENCES

D. ARMBRUSTER, J. GUCKENHEIMER, AND P. J. HOLMES [1987], *Heteroclinic cycles and modulated traveling waves in a system with $O(2)$ symmetry*, Phys. D, 29, pp. 257-282.

G. BREDON [1972], *Introduction to compact transformation groups*, in Pure and Applied Mathematics, Vol. 46, Academic Press, New York, London.

T. BRÖCKER AND T. TOM DIECK [1985], *Representations of Compact Lie Groups*, Graduate Texts in Mathematics, Springer-Verlag, New York.

E. BUZANO AND M. GOLUBITSKY [1983], *Bifurcations on the hexagonal lattice and the Bénard problem*, Philos. Trans. Roy. Soc. London Ser. A, 308, pp. 617-667.

P. CHOSSAT [1986], *Bifurcation secondaire de solution quasi périodiques dans un problème de bifurcation de Hopf par symétrie $O(2)$*, C. R. Acad. Sci. Paris Sér. I, 302, pp. 539-541.

P. CHOSSAT AND M. GOLUBITSKY [1988], *Iterates of maps with symmetry*, SIAM J. Math. Anal., 19, pp. 1259-1270.

E. N. DANCER [1980], *An implicit function theorem with symmetries and its application to nonlinear eigenvalue problems*, Bull. Austral. Math. Soc., 21, pp. 404-437.

G. DANGLEMAYR [1986], *Steady state mode interactions in the presence of $O(2)$ symmetry*, Dynamics and Stability of Systems, 1, pp. 159-185.

M. J. FIELD [1980], *Equivariant dynamical systems*, Trans. Amer. Math. Soc., 259, pp. 185-205.

——— [1988], *Equivariant bifurcation theory and symmetry breaking*, preprint, University of Sydney, Sydney, Australia.

M. GOLUBITSKY AND V. W. GUILLEMIN [1974], *Stable Mappings and Their Singularities*, Springer-Verlag, New York.

M. GOLUBITSKY AND D. SCHAEFFER [1985], *Singularities and Groups in Bifurcation Theory, Vol.* I, Applied Mathematical Sciences, Vol. 51, Springer-Verlag, New York.

M. GOLUBITSKY AND I. STEWART [1985], *Hopf bifurcation in the presence of symmetry*, Arch. Rational Mech. Anal., 87, pp. 107-165.

M. GOLUBITSKY, I. STEWART, AND D. SCHAEFFER [1988], *Singularities and Groups in Bifurcation Theory, Vol.* II, Applied Mathematical Sciences, Vol. 69, Springer-Verlag, New York.

V. GUILLEMIN AND A. POLLACK [1974], *Differential Topology*, Prentice-Hall, Englewood Cliffs, NJ.

G. IOOSS [1986], *Secondary bifurcations of the Taylor vortices into wavy inflow and outflow boundaries*, J. Fluid. Mech., 173, pp. 273-288.

I. G. KEVREKEDIS, B. NICOLAENCO, AND J. C. SCOVEL [1988], *Back in the saddle again: a computer assisted study of the Kuramoto-Sivashinsky equation*, SIAM J. Appl. Math., 50 (1990), pp. 760-790.

E. MOUTRANE [1988], *Interaction de modes sphériques dans le problème de Bénard entre deux sphères*, Ph.D. thesis, Université Nice, Nice, France.

V. POENARU [1976], *Singularités C^∞ en présence de symétrie*, Lecture Notes in Math., 510, Springer-Verlag, Berlin, New York.

D. RUELLE [1973], *Bifurcations in presence of a symmetry group*, Arch. Rational Mech. Anal., 51, pp. 131–172.

A. VANDERBAUWHEDE, M. KRUPA, AND M. GOLUBITSKY [1989], *Secondary bifurcations in symmetric systems*, in Proc. Equadiff Conference 1987, Lecture Notes in Pure and Appl. Math., Vol. 118, Marcel Dekker, New York, Basel, 1989.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1487-1510, November 1990
© 1990 Society for Industrial and Applied Mathematics
007

MATHEMATICAL ANALYSIS OF THE GUIDED MODES OF AN OPTICAL FIBER*

A. BAMBERGER† AND A. S. BONNET‡

Abstract. A mathematical formulation for the guided modes of an optical fiber is derived from Maxwell's equations: this formulation leads to an eigenvalue problem for a family of self-adjoint noncompact operators. The main spectral properties of these operators are established. Then the min-max principle provides an expression of the nonlinear dispersion relation, which connects the propagation constants of guided modes to the frequency. Various existence results are finally proved and a complete description of the dispersion curves (monotonicity, asymptotic behavior, existence of cutoff values) is carried out.

Key words. spectral analysis, guided modes, optical fiber

AMS(MOS) subject classifications. 35P, 78

0. Introduction. During the past 15 years, optical telecommunications and then integrated optics have undergone considerable development, which revived interest in the study of dielectrical guides.

Classically, the wave propagation in a cylindrical guide is based on the determination of the eigenmodes of the structure: these are electromagnetic waves of the form $\Phi(x)\, e^{i(\omega t-\beta z)}$, where x denotes the vector of transverse coordinates and z the longitudinal coordinate. The guided modes are of particular interest: they propagate without attenuation (β real) and the wave amplitude decays exponentially as the distance to the core of the guide increases.

The guided modes of a circular step-index fiber are well known (cf. [16], [17]). Indeed, in that case, the dispersion relation between the propagation constant β and the pulsation ω can be given explicitly by means of Bessel functions. The generic aspect of the dispersion curves for a circular step-index fiber is shown in Fig. 1.

Actually, in practice, there is a wide range of dielectrical guides (step-index fibers of arbitrary geometry, graded-index fibers, two cores fibers, $\cdots$) which cannot be treated analytically. Up until now, these cases have been studied by means of numerical computations (see, for example, [13], [18], [28], [29]). Only the one-dimensional problem of the planar waveguide has been considered in various theoretical studies (cf. [9], [27]).

In this work, we present a mathematical analysis of guided modes, for optical fibers having an arbitrary index profile in the core region.

Physical assumptions, equations, and notation are presented in § 1.

Then the first step consists in choosing between many mathematical formulations, resulting from Maxwell's equations. A previous study (cf. [3]) led us to a formulation on the magnetic field H, which we describe in § 2.

This formulation leads to an eigenvalue problem of the form $C_\beta H = \omega^2 H$, where C_β is an unbounded self-adjoint operator, with noncompact resolvent. Section 3 is concerned with a general study of the spectrum of the operator C_β. By deriving bounds for the eigenvalues, we establish a necessary condition for the existence of guided modes: the refraction index somewhere in the core region must be greater than the

* Received by the editors November 24, 1986; accepted for publication (in revised form) August 2, 1989.
† Institut Français du Pétrole, 1-4 avenue du Bois-Préau, BP 311, 92506 Rueil Malmaison Cedex, France.
‡ Groupe Hydrodynamique Navale, E.N.S.T.A. Centre de l'Yvette, Chemin de la Humière, 91120 Palaiseau, France.

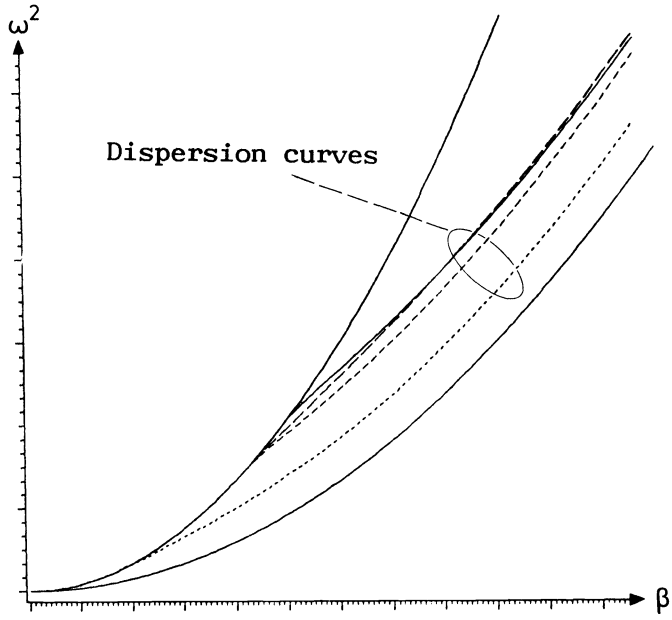


FIG. 1. *Dispersion curves of a circular step-index fiber.*

cladding index. Then after determining the essential spectrum, corresponding to a continuum of radiation modes, we apply the min-max principle to get an expression of the dispersion relation.

In § 4, we study the eigenvalues $\omega(\beta)$ as functions of β. A complete description of the dispersion curves (regularity, monotonicity, asymptotic behavior) is carried out and some existence results are derived. First we exhibit a category of index profiles such that the fiber supports at least two guided modes for every value of the propagation constant β. In the general case, we prove that the number of guided modes tends to infinity as β increases. This number varies at some special values of β which are called the cutoff values.

These cutoff values are studied in § 5. We prove that they are solutions of a nonlinear eigenvalue problem, set in a weighted Sobolev space, and that they form a sequence tending to infinity. For a given value of the propagation constant β, the fiber therefore admits at most a finite number of guided modes.

The results contained in this paper were announced in [2] and are part of the doctoral thesis [4] of A. S. Bonnet.

1. Modelling and notation.

1.1. The physical model. An optical fiber is a cylindrical structure which consists of a core of a dielectric material, surrounded by a cladding of another dielectric material (cf. Fig. 2). When the refractive index of the cladding is less than that of the core, the fiber is a waveguide. Electromagnetic waves can be propagated along the fiber without becoming deformed, their energy remaining confined in the core region.

We use a quite classical model (cf. [11], [16], [17], [20], [24], [25]): the fiber is assumed to be infinitely extended along its axis, denoted by Ox_3, and perfectly cylindrical (see [26] for the scattering theory of a deformed optical waveguide).

We suppose moreover that the cladding is infinitely extended in the transverse plane (O, x_1, x_2). This assumption is justified because the radius of the core is, in

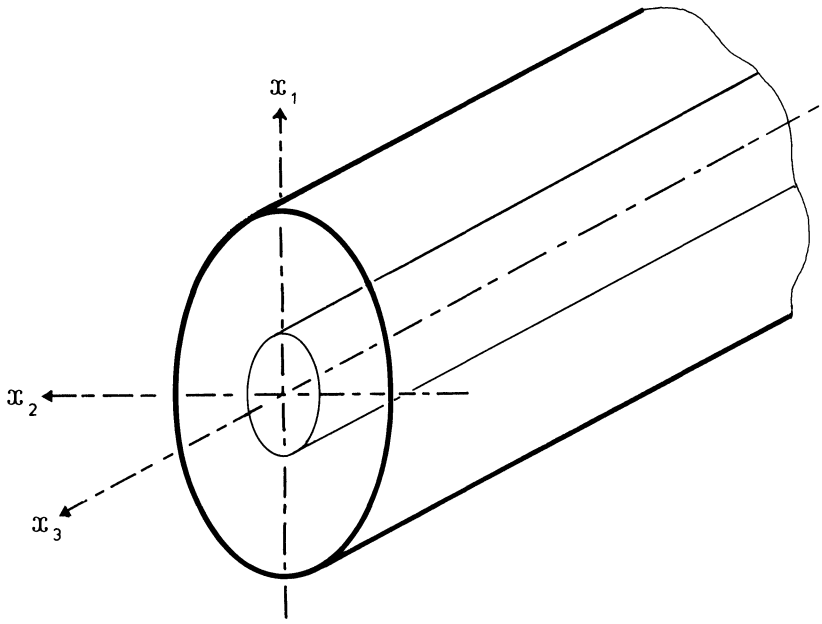


FIG. 2

practice, very small compared to the radius of the cladding, and because the guided modes have exponentially decaying fields in the cladding.

The fiber is completely determined by its index profile n, which is a bounded, positive function of the couple of transverse coordinates $x = (x_1, x_2)$. In the following, we just assume: $n \in L^\infty(\mathbb{R}^2)$ and $\inf\{n(x);\ x \in \mathbb{R}^2\} > 0$. Except for Proposition 3.1, it is not necessary for the index profile to be more regular.

In the following, the fiber cladding is supposed to be homogeneous (see [4] and [5] for some generalizations). More precisely, we suppose that there is a refractive index n_∞ and a bounded domain Ω such that $n(x) = n_\infty$ if $x \notin \Omega$. We say that Ω is the *core region*, the exterior domain $\mathbb{R}^2/\Omega$ is the *cladding*, and n_∞ is the refractive index of the cladding. We will also denote by D_a a disk of centre O and radius a which contains the core region Ω.

Some particular categories of fibers which are used in practice will be mentioned as references:

(a) A fiber whose index profile n is piecewise constant is called a step-index fiber. A step-index fiber is said to be circular (respectively, elliptic) if there is a circular (respectively, elliptic) domain Ω such that the index profile is constant in Ω and outside Ω.

(b) A fiber is called a graded-index fiber when its index profile n belongs to $\mathscr{C}^2(\mathbb{R}^2)$.

1.2. The equations for the guided modes. The cylindrical geometry of the fiber suggests that we look for particular solutions of the Maxwell equations

$$\text{Rot}\,\mathbb{H} = \varepsilon_0 n^2 \frac{\partial \mathbb{E}}{\partial t}, \qquad \text{Rot}\,\mathbb{E} = -\mu_0 \frac{\partial \mathbb{H}}{\partial t}, \tag{1.1}$$

which can be written as

$$\begin{pmatrix} \mathbb{E} \\ \mathbb{H} \end{pmatrix}(x_1, x_2, x_3, t) = \begin{pmatrix} E \\ H \end{pmatrix}(x_1, x_2)\, e^{i(kc_0 t - \beta x_3)}, \tag{1.2}$$

where $\mathbb{E}$ is the electric field, $\mathbb{H}$ the magnetic field, c_0 the velocity of light in the vacuum, and ε_0 and μ_0 the dielectric permittivity and the magnetic permeability of the vacuum, respectively. The variables k and β are called, respectively, the *wavenumber* and the *propagation constant.*

Such a solution (E, H, β, k) is said to be a *guided mode* if, moreover,

$$(\beta, k) \in \mathbb{R}^2, \quad (E, H) \neq (0, 0) \quad \text{and} \quad (E, H) \in [L^2(\mathbb{R}^2)]^3 \times [L^2(\mathbb{R}^2)]^3.$$

For fields of the form (1.2), system (1.1) becomes

$$\text{Rot}_\beta H = i\varepsilon_0 c_0 k n^2 E, \qquad \text{Rot}_\beta E = -i\mu_0 c_0 k H. \tag{1.3}$$

The index β associated with Rot operator means that the derivation with respect to x_3 is replaced by the multiplication by $(-i\beta)$.

We are interested in the description of the set of all pairs (β, k) in $\mathbb{R}^2$ such that system (1.3) has nontrivial square integrable solutions. Our aim is therefore to establish and study *the dispersion relation* between β and k.

Note that for every fixed β, we get a *two-dimensional eigenvalue problem* where the wavenumber k is the eigenvalue and the electromagnetic field (E, H) the associated eigenvector.

Suppose that (E, H, β, k) is a guided mode. Then we can easily show that $(E, -H, \beta, -k)$, $(E', H', -\beta, k)$, and $(E', -H', -\beta, -k)$ are guided modes as well, where E' and H' denote the symmetrics of E and H with respect to the transverse plane (O, x_1, x_2). Thus we can restrict ourselves in the following to the pairs (β, k) of positive real numbers.

We first need to introduce some notation.

1.3. Notation. Denote by $\mathcal{D}(\mathbb{R}^2)$ the space of indefinitely differentiable complex-valued functions which have compact support in $\mathbb{R}^2$, and by $\mathcal{D}'(\mathbb{R}^2)$ the space of complex-valued distributions on $\mathbb{R}^2$.

Let $\varphi \in \mathcal{D}'(\mathbb{R}^2)$ and

$$F = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix} \in [\mathcal{D}'(\mathbb{R}^2)]^3.$$

We define

$$\text{grad } \varphi = \begin{bmatrix} \dfrac{\partial \varphi}{\partial x_1} \\ \dfrac{\partial \varphi}{\partial x_2} \end{bmatrix}, \qquad \text{rot } \varphi = \begin{bmatrix} \dfrac{\partial \varphi}{\partial x_2} \\ -\dfrac{\partial \varphi}{\partial x_1} \end{bmatrix},$$

$$\Delta \varphi = \frac{\partial^2 \varphi}{\partial x_1^2} + \frac{\partial^2 \varphi}{\partial x_2^2}, \qquad \mathbf{F} = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix},$$

$$\text{rot } \mathbf{F} = \frac{\partial F_2}{\partial x_1} - \frac{\partial F_1}{\partial x_2}, \qquad \text{div } \mathbf{F} = \frac{\partial F_1}{\partial x_1} + \frac{\partial F_2}{\partial x_2},$$

and for $\beta \in \mathbb{R}$

$$\text{Div}_\beta F = \frac{\partial F_1}{\partial x_1} + \frac{\partial F_2}{\partial x_2} - i\beta F_3, \qquad \text{Rot}_\beta F = \begin{bmatrix} \dfrac{\partial F_3}{\partial x_2} + i\beta F_2 \\ -i\beta F_1 - \dfrac{\partial F_3}{\partial x_1} \\ \dfrac{\partial F_2}{\partial x_1} - \dfrac{\partial F_1}{\partial x_2} \end{bmatrix},$$

$$\mathrm{Grad}_\beta \varphi = \begin{bmatrix} \dfrac{\partial \varphi}{\partial x_1} \\ \dfrac{\partial \varphi}{\partial x_2} \\ -i\beta\varphi \end{bmatrix}.$$

The following identities hold:

$$\mathrm{rot}\,(\mathrm{rot}\,\mathbf{F}) = -\Delta\mathbf{F} + \mathrm{grad}\,(\mathrm{div}\,\mathbf{F}),$$

$$\mathrm{Rot}_\beta\,(\mathrm{Rot}_\beta F) = -\Delta F + \beta^2 F + \mathrm{Grad}_\beta\,(\mathrm{Div}_\beta F).$$

The inner product and norm in $[L^2(\mathbb{R}^2)]^j$ $(j=1,2,3)$ will be denoted, without distinction, by $(\cdot,\cdot)$ and $|\cdot|_2$.

2. The choice of the variational formulation.

2.1. Classical formulations. Many mathematical formulations have already been established for problems of electromagnetics (cf. [3] and [7]). We briefly recall some of them, dwelling on the reasons which have motivated our choice.

At first, note that we can derive from system (1.3) some equivalent equations by eliminating either E or H.

LEMMA 2.1. *Let* (E, H) *be a solution of system* (1.3). *Then*

$$\mathrm{Rot}_\beta\,(\mathrm{Rot}_\beta E) = k^2 n^2 E, \tag{2.1}$$

$$\mathrm{Div}_\beta\,(n^2 E) = 0, \tag{2.2}$$

$$\mathrm{Rot}_\beta\left(\frac{1}{n^2}\,\mathrm{Rot}_\beta H\right) = k^2 H, \tag{2.3}$$

$$\mathrm{Div}_\beta H = 0. \tag{2.4}$$

Proof. Equations (2.1) and (2.3) are obtained by applying the operator Rot_β to (1.3). Equations (2.2) and (2.4) are deduced from (2.1) and (2.3) using the following identity:

$$\mathrm{Div}_\beta\,(\mathrm{Rot}_\beta F) = 0 \quad \forall F \in [\mathscr{D}(\mathbb{R}^2)]^3. \tag{2.5}$$

□

By (2.2), the electric field E does not belong to the space $[H^1(\mathbb{R}^2)]^3$ if the index profile n is not regular. Hence, a formulation on the magnetic field is more convenient: indeed, H belongs to the space $[H^1(\mathbb{R}^2)]^3$ for every index profile (see § 2.2).

Using (2.4), it is possible to eliminate the longitudinal component H_3 in system (2.3). The derived formulation involves only H_1 and H_2 as unknowns, but is no longer symmetrical. For that reason, we prefer keeping the three unknowns H_1, H_2, and H_3.

The natural variational formulation associated with (2.3) uses the following bilinear form:

$$b(\beta; H, H') = \int_{\mathbb{R}^2} \frac{1}{n^2}\,\mathrm{Rot}_\beta H \cdot \overline{\mathrm{Rot}_\beta H'}\, dx,$$

which is not elliptic on the space $[H^1(\mathbb{R}^2)]^3$, but on the subspace

$$\{H \in [H^1(\mathbb{R}^2)]^3;\ \mathrm{Div}_\beta H = 0\}.$$

This space depends on β, which is an unknown and may vary. To obviate this difficulty, a modification used for solving Maxwell's equations in bounded domains (cf. [14]) is to introduce the following equation:

$$\text{Rot}_\beta\left(\frac{1}{n^2}\,\text{Rot}_\beta H\right) - s\,\text{Grad}_\beta\,(\text{Div}_\beta H) = k^2 H, \tag{2.6}$$

where s is some arbitrary positive real.

Every solution of (2.3) satisfies (2.6). The converse will be established in the next section. Moreover, the bilinear form associated with (2.6) is elliptic on $(H^1(\mathbb{R}^2))^3$.

2.2. A formulation for the magnetic field. Let us denote by V the Sobolev space $(H^1(\mathbb{R}^2))^3$ equipped with the usual norm $\|F\|_V = [|F|_2^2 + |\text{grad } F|_2^2]^{1/2}$, and by $V(\beta)$ the space of all square integrable fields F, such that $\text{Rot}_\beta F$ and $\text{Div}_\beta F$ are also square integrable. Equipped with the norm $\|F\|_{V(\beta)} = [|F|_2^2 + |\text{Rot}_\beta F|_2^2 + |\text{Div}_\beta F|_2^2]^{1/2}$, $V(\beta)$ is a Hilbert space (cf. [7]).

We first prove that the magnetic field belongs to V.

LEMMA 2.2. (i) *Let* (E, H) *be the solution of system* (3.1). *Then* $(E, H) \in (L^2(\mathbb{R}^2))^3 \times (L^2(\mathbb{R}^2))^3$ *if and only if* $H \in V(\beta)$.

(ii) *For every* F *in* V

$$\int_{\mathbb{R}^2} \{|\text{Rot}_\beta F|^2 + |\text{Div}_\beta F|^2\}\,dx = \int_{\mathbb{R}^2} \{|\text{grad } F|^2 + \beta^2 |F|^2\}\,dx. \tag{2.7}$$

(iii) *Hilbert spaces* V *and* $V(\beta)$ *are isomorphic.*

Proof. Assertion (i) is a trivial consequence of (2.3) and (2.4).

(ii) For every F in $[\mathcal{D}(\mathbb{R}^2)]^3$:

$$\begin{aligned}\int_{\mathbb{R}^2} |\text{Rot}_\beta F|^2\,dx &= \int_{\mathbb{R}^2} \text{Rot}_\beta\,(\text{Rot}_\beta F)\cdot \bar{F}\,dx \\ &= \int_{\mathbb{R}^2} \{-\Delta F + \beta^2 F + \text{Grad}_\beta\,(\text{Div}_\beta F)\}\cdot \bar{F}\,dx \\ &= \int_{\mathbb{R}^2} \{|\text{grad } F|^2 + \beta^2 |F|^2 - |\text{Div}_\beta F|^2\}\,dx.\end{aligned}$$

Since $(\mathcal{D}(\mathbb{R}^2))^3$ is dense in V, this equality holds for every F in V.

(iii) By [7], the space $(\mathcal{D}(\mathbb{R}^2))^3$ is also dense in $V(\beta)$. Hence assertion (iii) is a consequence of identity (2.7). □

Remarks 2.1. (1) Identity (2.7) is analogous to the following classical one:

$$\int_{\mathbb{R}^2} \{|\text{rot } \mathbf{F}|^2 + |\text{div } \mathbf{F}|^2\}\,dx = \int_{\mathbb{R}^2} |\text{grad } \mathbf{F}|^2\,dx. \tag{2.8}$$

(2) Note that the functional space V neither depends on β nor on the index profile n. □

We can now establish the following theorem.

THEOREM 2.1. *The following assertions are equivalent*:

(i) (E, H) *is a nontrivial solution of* (1.3) *and* $E, H \in [L^2(\mathbb{R}^2)]^3$.

(ii) H *is a nontrivial solution of* (2.6) *and* $H \in V$.

Proof. By Lemmas 2.1 and 2.2, (ii) is a consequence of (i). Conversely, let H be a solution of (2.6). We first need to prove that H satisfies (2.3). By taking the divergence

of (2.6), we get

$$-s \operatorname{Div}_\beta (\operatorname{Grad}_\beta \varphi) = k^2 \varphi,$$

where $\varphi = \operatorname{Div}_\beta H$. Thus φ satisfies $-s\Delta\varphi = (k^2 - \beta^2 s)\varphi$ and $\varphi \in L^2(\mathbb{R}^2)$. Consequently, φ must vanish everywhere and (2.3) holds.

Moreover, by (2.7), the wavenumber k cannot be equal to zero since H does not vanish identically. Therefore we can set

$$E = \frac{1}{i\varepsilon_0 c_0 k n^2} \operatorname{Rot}_\beta H,$$

and (E, H) is a nontrivial solution of (1.3). □

In order to choose the value of s, we note that in the outside region, with refractive index n_∞, equation (2.6) becomes

$$-\frac{1}{n_\infty^2}\Delta H + \frac{\beta^2}{n_\infty^2} H + \left(\frac{1}{n_\infty^2} - s\right) \operatorname{Grad}_\beta (\operatorname{Div}_\beta H) = k^2 H. \tag{2.9}$$

By taking $s = 1/n_\infty^2$, we get uncoupled Helmholtz equations in this domain. The convenience of this choice will be confirmed in the following section.

2.3. Variational formulation. To conclude this section, let us prove that equation (2.9) leads to a variational formulation with a V-elliptic bilinear form.

Hereafter, we set

$$c(\beta; H, H') = \int_{\mathbb{R}^2} \left\{ \frac{1}{n^2} \operatorname{Rot}_\beta H \cdot \overline{\operatorname{Rot}_\beta H'} + \frac{1}{n_\infty^2} \operatorname{Div}_\beta H \, \overline{\operatorname{Div}_\beta H'} \right\} dx. \tag{2.10}$$

By Theorem 2.1, we can now consider the following problem:

(2.11) $(\mathcal{P})$

Find all pairs of reals (β, k) such that there exists H satisfying

$$H \in V, \qquad H \neq 0, \tag{2.12}$$
$$c(\beta; H, H') = k^2 (H, H') \quad \forall H' \in V.$$

The following lemma is a straightforward consequence of identity (2.7).

LEMMA 2.3. *The form $c(\beta)$ is Hermitian and satisfies, for every $H \in V$*

$$c(\beta; H, H) \geqq \frac{1}{n_+^2} \int_{\mathbb{R}^2} \{|\operatorname{grad} H|^2 + \beta^2 |H|^2\} \, dx, \tag{2.13}$$

$$c(\beta; H, H) \leqq \frac{1}{n_-^2} \int_{\mathbb{R}^2} \{|\operatorname{grad} H|^2 + \beta^2 |H|^2\} \, dx, \tag{2.14}$$

where n_+ and n_- are defined by

$$n_+ = \sup_{x \in \mathbb{R}^2} n(x) \quad \text{and} \quad n_- = \inf_{x \in \mathbb{R}^2} n(x). \tag{2.15}$$

In particular, $c(\beta)$ is V-elliptic for nonzero values of β.

3. Derivation of the dispersion relation. The problem $(\mathcal{P})$, defined by (2.11), can be written equivalently as follows:

(3.1) $(\mathcal{P})$

> Find all pairs $(\beta, k) \in (\mathbb{R}^+)^2$ such that there exists H satisfying
>
> $$H \in D(C_\beta), \quad H \neq 0, \quad C_\beta H = k^2 H,$$

where we denote by C_β the unbounded operator of $(L^2(\mathbb{R}^2))^3$, with domain $D(C_\beta)$, associated with the form $c(\beta)$. In other words, k^2 is an eigenvalue of C_β and H an associated eigenfield.

We therefore must study the spectrum $\sigma(\beta)$ of C_β and especially the set of its eigenvalues $\sigma_p(\beta)$.

In this section we first derive a lower bound for $\sigma(\beta)$ and an upper bound for $\sigma_p(\beta)$. The essential spectrum $\sigma_{\text{ess}}(\beta)$ is then determined. The dispersion relation between β and k is finally derived by using the min-max principle.

3.1. Lower and upper bounds for the eigenvalues. Using Lemma 2.3, we first establish Lemma 3.1.

LEMMA 3.1. *For every nonnegative real β, the operator C_β is self-adjoint. Moreover,*

(i) $\sigma(\beta) \subset [\beta^2/n_+^2, +\infty[$,

(ii) $\sigma_p(\beta) \subset]\beta^2/n_+^2, +\infty[$,

where n_+ is defined by (2.15).

Proof. The self-adjointness of C_β and inclusion (i) are consequences of (2.13) (cf. [23]).

Suppose now that $c(\beta; H, H) = (\beta/n_+)^2 |H|_2^2$ for some H in V. It therefore follows from (2.13) that

$$\int_{\mathbb{R}^2} |\text{grad } H|^2 \, dx = 0,$$

which implies that H is constant. Since H belongs to $[L^2(\mathbb{R}^2)]^3$, it must then vanish everywhere. Consequently, $(\beta/n_+)^2$ cannot be an eigenvalue of C_β. □

Remark 3.1. By the previous lemma, every solution (β, k) of $(\mathcal{P})$ must satisfy

$$k^2 > \left(\frac{\beta}{n_+}\right)^2.$$

This means that the guided wave always propagates faster than a plane wave in an homogeneous medium of index n_+.

To derive an upper bound for the eigenvalues, the following additional hypothesis is required:

(3.2) There exists a finite collection $\{\Omega_1, \cdots, \Omega_m\}$ of open regular subsets of $\mathbb{R}^2$ such that

$$\mathbb{R}^2 = \bigcup_{j=1}^{m} \bar{\Omega}_j \quad \text{and} \quad n|_{\bar{\Omega}_j} \in \mathcal{C}^2(\bar{\Omega}_j) \quad \text{for } j = 1, \cdots, m.$$

In other words, the index profile is supposed to be piecewise regular. This hypothesis is sufficient but probably not necessary. However, assumption (3.2) is satisfied by all fibers used in practice (step-index fibers, graded-index fibers, two-fiber couplers).

PROPOSITION 3.1. *Assume that the index profile n satisfies* (3.2). *Then* $\sigma_p(\beta) \subset]-\infty, \beta^2/n_\infty^2]$.

Proof. (1) Let $H \in V$ and $\lambda > (\beta/n_\infty)^2$ such that $C_\beta H = \lambda H$. Then H satisfies $-\Delta H + (\lambda n_\infty^2 - \beta^2)H = 0$ outside the core region Ω. Hence we deduce from the uniqueness result proved in [21] that H vanishes in $\mathbb{R}^2 \setminus D_a$, where D_a is a disk containing the core region Ω.

(2) Under assumption (3.2), H must vanish identically (cf. [10, pp. 190–192]). □

Remarks 3.2. (1) Notice that condition (3.2) has been used only for the second part of the proof, which requires a continuation principle.

(2) By the previous lemma, every solution (β, k) of problem $(\mathcal{P})$, defined by (3.1), must satisfy the following inequality:

$$k^2 \leqq \left(\frac{\beta}{n_\infty}\right)^2.$$

This means that the guided wave always propagates slower than a plane wave in a homogeneous medium of index n_∞.

These first results are summarized in Corollary 3.2.

COROLLARY 3.2. *Assume that* (3.2) *holds*:

(i) *If* $n_+ = n_\infty$ *or* $\beta = 0$, *then* $\sigma_p(\beta) = \varnothing$.

(ii) *If* $n_+ > n_\infty$ *and* $\beta \neq 0$, *then* $\sigma_p(\beta) \subset]\beta^2/n_+^2, \beta^2/n_\infty^2]$.

In other words, the problem $(\mathcal{P})$, defined by (3.1), has no solution when n_+ is equal to n_∞. For this reason, we shall assume in the sequel that the index profile always satisfies the "guidance condition":

$$\boxed{n_+ > n_\infty}.$$

3.2. The essential spectrum. The aim of this section is to establish the following proposition.

PROPOSITION 3.2. *For every nonnegative real* β, *the essential spectrum of* C_β *is given by*

$$\sigma_{\text{ess}}(\beta) = \left[\frac{\beta^2}{n_\infty^2}, +\infty\right[.$$

Remark 3.3. The essential spectrum corresponds to the continuum of propagation constants of radiation modes. See, for example, [16] for a description of the radiation modes of a cylindrical step-index fiber.

To prove this proposition, we shall use another version of the bilinear form $c(\beta)$. Let us define

$$d_0(H, H') = \int_{\mathbb{R}^2} \left\{ \frac{1}{n^2} \operatorname{rot} \mathbf{H} \operatorname{rot} \bar{\mathbf{H}}' + \frac{1}{n_\infty^2} \operatorname{div} \mathbf{H} \operatorname{div} \bar{\mathbf{H}}' + \frac{1}{n^2} \operatorname{grad} H_3 \cdot \operatorname{grad} \bar{H}_3' \right\} dx, \tag{3.3}$$

$$d_1(H, H') = i \int_{\mathbb{R}^2} \left(\frac{1}{n^2} - \frac{1}{n_\infty^2}\right) \{\operatorname{grad} H_3 \cdot \bar{\mathbf{H}}' - \mathbf{H} \cdot \operatorname{grad} \bar{H}_3'\} \, dx, \tag{3.4}$$

$$d_2(H, H') = \int_{\mathbb{R}^2} \left(\frac{1}{n^2} - \frac{1}{n_\infty^2}\right) \mathbf{H} \cdot \bar{\mathbf{H}}' \, dx, \tag{3.5}$$

$$d(\beta; H, H') = d_0(H, H') + \beta d_1(H, H') + \beta^2 d_2(H, H'). \tag{3.6}$$

Then we have the following lemma.

LEMMA 3.3. (i) *The bilinear form* $c(\beta)$ *admits the following expression*:

$$c(\beta; H, H') = d(\beta; H, H') + \frac{\beta^2}{n_\infty^2}(H, H'). \tag{3.7}$$

(ii) *The forms* d_0, d_1, *and* d_2 *are Hermitian and continuous on* V.

(iii) *The form* d_0 *satisfies*

$$\forall H \in V, \quad d_0(H, H) \geqq \frac{1}{n_+^2} \int_{\mathbb{R}^2} |\text{grad } H|^2 \, dx. \tag{3.8}$$

(iv) *The forms* d_1 *and* d_2 *are compact perturbations of* d_0.

Proof. The decomposition (i) follows easily from the Green formula:

$$\int_{\mathbb{R}^2} \{\text{grad } H_3 \cdot \bar{\mathbf{H}}' - \mathbf{H} \cdot \text{grad } \bar{H}_3'\} \, dx$$

$$= \int_{\mathbb{R}^2} \{(\text{div } \mathbf{H})\bar{H}_3' - H_3(\text{div } \bar{\mathbf{H}}')\} \, dx.$$

To prove (ii), we first note that

$$d_0(H, H) \geqq \frac{1}{n_+^2} \int_{\mathbb{R}^2} \{|\text{grad } H_3|^2 + |\text{rot } \mathbf{H}|^2 + |\text{div } \mathbf{H}|^2\} \, dx$$

and then we use equality (2.8).

Now, since $1/n(x)^2 - 1/n_\infty^2 = 0$ outside of Ω, (iv) becomes a straightforward consequence of the compact injection of $H^1(\Omega)$ into $L^2(\Omega)$. □

We are now ready for the proof of Proposition 3.2.

Proof of Proposition 3.2. Let us denote by D_β the unbounded operator associated with the bilinear form $d(\beta; H, H)$ defined by (3.6) and by $\sigma_{\text{ess}}(D_\beta)$ its essential spectrum. By (3.7), we must establish the following identity:

$$\sigma_{\text{ess}}(D_\beta) = [0, +\infty[. \tag{3.9}$$

Now the proof of (3.9) involves two parts:

— First the essential spectrum of D_β is proved to be independent of β. Indeed, by the previous lemma the forms d_1 and d_2 are compact perturbations of d_0. By the Weyl theorem (cf. [23]), they do not modify the essential spectrum. In other words, $\sigma_{\text{ess}}(D_\beta) = \sigma_{\text{ess}}(D_0)$ for every value of β.

— It remains to prove that $\sigma_{\text{ess}}(D_0) = [0, +\infty[$.

By Lemma 3.1, the following inclusion holds: $\sigma_{\text{ess}}(D_0) \subset [0, +\infty[$, and we must just establish the converse inclusion. Moreover, $\sigma_{\text{ess}}(D_0)$ is closed. Hence we must actually prove that $]0, +\infty[\subset \sigma_{\text{ess}}(D_0)$.

We do it by using singular sequences (cf. [23]). Let γ be a strictly positive real. Let ψ be some function of $\mathcal{D}(\mathbb{R}^2)$ which vanishes in Ω, let $H^{(0)}$ be some arbitrary vector of $\mathbb{C}^3$, and let J_0 denote the Bessel function of first kind of order zero (cf. [1]). We define a sequence $(H^{(p)})$ as follows:

$$H^{(p)} = \frac{1}{\sqrt{p}} \psi\left(\frac{x}{p}\right) J_0(\sqrt{\gamma}\, x) H^{(0)}.$$

To prove that $(H^{(p)})$ is a singular sequence, we must establish the three following statements:

$$|H^{(p)}|_2 \geqq \alpha > 0 \quad \forall p \in \mathbb{N}, \tag{3.10a}$$

$$|D_0 H^{(p)} - \gamma H^{(p)}|_2 \xrightarrow[p\to+\infty]{} 0, \tag{3.10b}$$

$$H^{(p)} \xrightarrow[p\to+\infty]{} 0 \quad \text{for } (L^2(\mathbb{R}^2))^3, \text{ weakly.} \tag{3.10c}$$

Assertions (3.10a) and (3.10b) can be deduced from the asymptotic behavior of J_0 (cf. [1]), and (3.10c) is a direct consequence of the Lebesgue dominated convergence theorem. □

3.3. The min-max principle and the dispersion relation. By Propositions 3.1 and 3.2, if (3.2) holds, all eigenvalues of C_β—except perhaps $(\beta/n_\infty)^2$—are located below the essential spectrum. Hence we can apply the min-max principle to get expressions of the eigenvalues as functions of β.

Let us define

$$\lambda_1(\beta) = \inf_{\substack{H\in(L^2(\mathbb{R}^2))^3 \\ |H|_2=1}} c(\beta; H, H)$$

and if $m>1$

$$\lambda_m(\beta) = \sup_{H^{(1)},\cdots,H^{(m-1)}\in(L^2(\mathbb{R}^2))^3} \ \inf_{\substack{H\in[H^{(1)},\cdots,H^{(m-1)}]_V^\perp \\ |H|_2=1}} c(\beta; H, H), \tag{3.11}$$

where $[H^{(1)}, \cdots, H^{(m-1)}]_V^\perp = \{H \in V;\ (H, H^{(j)}) = 0;\ j = 1, m-1\}$.

By the min-max principle (cf. [22]), we have, for each fixed m

$$(\beta/n_+)^2 < \lambda_1(\beta) \leqq \lambda_2(\beta) \cdots \leqq \lambda_m(\beta) \leqq (\beta/n_\infty)^2 \tag{3.12}$$

and EITHER

(i) $\lambda_m(\beta) < (\beta/n_\infty)^2$.

In that case, there are m eigenvalues of C_β (counting them a number of times equal to their multiplicity) below $(\beta/n_\infty)^2$, and $\lambda_m(\beta)$ is the mth eigenvalue.

OR

(ii) $\lambda_m(\beta) = (\beta/n_\infty)^2$.

In that case,

$$\lambda_m(\beta) = \lambda_{m+1}(\beta) = \lambda_{m+2}(\beta) = \cdots = (\beta/n_\infty)^2$$

and there are at most $(m-1)$ eigenvalues of C_β (counting multiplicity) below $(\beta/n_\infty)^2$.

Note particularly that if (3.2) holds all eigenvalues of C_β—except perhaps $(\beta/n_\infty)^2$—have finite multiplicity and are isolated points of $\sigma(\beta)$.

It is now clear that the min-max principle provides an expression of the dispersion relation between the wavenumber k and the propagation constant β as follows.

THEOREM 3.1. *The solutions* (β, k) *of problem* $(\mathcal{P})$, *defined by* (3.1), *such that* $k < \beta/n_\infty$ *are the roots of the dispersion relation*:

$$\boxed{\begin{aligned} &\lambda_m(\beta) = k^2, \\ &\lambda_m(\beta) < (\beta/n_\infty)^2, \end{aligned} \qquad m = 1, 2, \cdots.} \tag{3.13}$$

Remarks 3.4. (1) The min-max principle does not allow us to take into account the solutions (β, k) of problem $(\mathcal{P})$ such that $k = \beta/n_\infty$. These solutions are different

from the others: indeed the associated field does not decrease exponentially at infinity (it satisfies $\Delta H = 0$ outside the core region). From a physical viewpoint, the corresponding mode is not "guided." However, the case $k = \beta/n_\infty$ will be considered in § 4 for the study of cutoff values.

(2) The min-max principle is useful for comparing eigenvalues for various indices. Indeed, let n and n' be two index profiles such that

$$n(x) \leqq n'(x) \quad \text{a.e. in } \mathbb{R}^2. \tag{3.14}$$

Then $c(\beta, n'; H, H) \leqq c(\beta, n; H, H)$ for every H and every β. Consequently,

$$\lambda_m(\beta, n') \leqq \lambda_m(\beta, n) \quad \forall \beta > 0, \quad m = 1, 2, \cdots.$$

This obvious consequence of the min-max principle has many practical applications. Indeed, for every index profile n, a circular step-index profile n' can be defined by

$$n'(x) = n_+ \quad \text{a.e. in } D_a,$$
$$n'(x) = n_\infty \quad \text{a.e. in } \mathbb{R}^2 \backslash D_a$$

such that (3.14) is satisfied. Since the functions $\lambda_m(\beta, n')$ are well known, this comparison principle immediately provides information about the functions $\lambda_m(\beta, n)$. □

4. Study of the dispersion relation (3.13).

4.1. Differentiability and monotonicity results for the dispersion curves. The aim of this section is to establish Proposition 4.1.

PROPOSITION 4.1. (i) *The functions $\beta \to \lambda_m(\beta)$ are continuous and almost everywhere differentiable for $\beta \in \mathbb{R}^+$.*

(ii) *Suppose that*

$$\left(\frac{n_+}{n_\infty} + 1\right) \Delta(n) < \frac{1}{n_\infty^2}, \tag{4.1}$$

where

$$\Delta(n) = \sup_{x \in \mathbb{R}^2} \left| \frac{1}{n(x)^2} - \frac{1}{n_\infty^2} \right|. \tag{4.2}$$

Then the functions $\lambda_m(\beta)$ are strictly increasing in β for $\beta \in \mathbb{R}^+$.

This proposition is a straightforward consequence of the following lemma, which provides an estimate for the derivatives of the functions $\lambda_m(\beta)$.

LEMMA 4.1. *The function*

$$\Lambda_m(\beta) = \lambda_m(\beta) - (\beta/n_\infty)^2 \tag{4.3}$$

is continuous. It is differentiable almost everywhere and its derivative satisfies

$$\left| \frac{d\Lambda_m}{d\beta}(\beta) \right| \leqq 2\beta \left(\frac{n_+}{n_\infty} + 1\right) \Delta(n), \tag{4.4}$$

where $\Delta(n)$ is defined by (4.2).

Proof. Note first that $\Lambda_m(\beta)$ is the mth max-min associated with the form $d(\beta; H, H)$ defined by (3.6).

Let β and β' be two distinct positive reals and let $H \in V$ such that $|H|_2 = 1$. Then

$$d(\beta; H, H) - d(\beta'; H, H) = (\beta - \beta') d_1(H, H) + (\beta^2 - \beta'^2) d_2(H, H), \tag{4.5}$$

where d_1 and d_2 are defined by (3.4) and (3.5).

By the Cauchy-Schwarz inequality

$$\begin{aligned} &|d_1(H, H)| \leqq 2\Delta(n)|\text{grad } H_3|_2^2, \\ &|d_2(H, H)| \leqq \Delta(n), \end{aligned} \tag{4.6}$$

and by (2.13)

$$\int_{\mathbb{R}^2} |\text{grad } H_3|^2 \, dx \leqq n_+^2 (d(\beta'; H, H) + (\beta'/n_\infty)^2). \tag{4.7}$$

Using (4.5)-(4.7), we obtain

$$\begin{aligned} d(\beta; H, H) \leqq\ & d(\beta'; H, H) + |\beta - \beta'|\Delta(n) \\ & \cdot \left\{ 2n_+ \left(d(\beta'; H, H) + \frac{\beta'^2}{n_\infty^2} \right)^{1/2} + (\beta + \beta') \right\}. \end{aligned} \tag{4.8}$$

Therefore, for every integer m

$$\begin{aligned} \Lambda_m(\beta) \leqq\ & \Lambda_m(\beta') + |\beta - \beta'|\Delta(n) \\ & \cdot \{2n_+(\Lambda_m(\beta') + (\beta'/n_\infty)^2)^{1/2} + (\beta + \beta')\}. \end{aligned}$$

By (3.12), $\Lambda_m(\beta')$ is always negative, so that we have

$$\Lambda_m(\beta) \leqq \Lambda_m(\beta') + |\beta - \beta'|\Delta(n) \left\{ 2\frac{n_+}{n_\infty}\beta' + (\beta + \beta') \right\}$$

and, by inverting β and β'

$$|\Lambda_m(\beta') - \Lambda_m(\beta)| \leqq |\beta - \beta'| 2\Delta(n) \max(\beta, \beta') \left\{ \frac{n_+}{n_\infty} + 1 \right\}.$$

Consequently, the function $\Lambda_m(\beta)$ is locally Lipschitz and hence almost everywhere differentiable. By the previous inequality, its derivative satisfies (4.4). □

Remarks 4.1. (i) Note that Lemma 4.1 cannot be improved, since, in the case of the circular step-index fiber, the functions $\lambda_m(\beta)$ are not everywhere differentiable. In fact, we can show that the eigenvalues are analytic functions of the propagation constant β (cf. [12]). However, since the analytic curves may intersect, the functions $\lambda_m(\beta)$ are just piecewise analytic.

(ii) By Theorem 4.1, if (4.1) holds, the functions $\lambda_m(\beta)$ are one to one and the dispersion relation (3.13) equivalently reads:

$$\begin{aligned} &\lambda_m^{-1}(k^2) = \beta, \\ &\lambda_m^{-1}(k^2) < kn_\infty, \end{aligned} \qquad m = 1, 2, \cdots.$$

Consequently, all the results concerning the direct problem, which consists in studying the solutions k for a given value of β, may be transposed to the inverse problem.

(iii) A simple calculation shows that if

$$n_+/n_\infty < Q^+ \quad \text{and} \quad n_\infty/n_- < Q^-, \tag{4.9}$$

where Q^+ is the greatest root of the equation $x^3 - x - 1 = 0$ and $Q^- = (1/(1+Q^+)+1)^{1/2}$, then the functions $\lambda_m(\beta)$ are strictly increasing in β for $\beta \in \mathbb{R}^+$.

The computation gives the following approximate values:

$$Q^+ \simeq 1.33, \qquad Q^- \simeq 1.20$$

and condition (4.9) is satisfied by all the fibers used in practice. □

4.2. Existence of eigenvalues: a sufficient condition. Let us recall that (see, e.g., [17]) a circular or elliptic step-index fiber has, for every nonzero value of β, at least two guided modes.

In the circular case $\lambda_1(\beta)=\lambda_2(\beta)<(\beta/n_\infty)^2$, and in the elliptic case $\lambda_1(\beta)<\lambda_2(\beta)<(\beta/n_\infty)^2$, for every nonzero value of the propagation constant β.

However, it may happen if there are some regions in D_a where the index profile n is lower than n_∞, that the fiber supports no guided modes for small values of β. This is, for example, the case of the so-called "W-profiles" (cf. [11], [20]).

The following theorem provides a sufficient condition on the index profile which ensures the existence of two guided modes for every nonzero value of β.

THEOREM 4.1. *Suppose*

$$\int_{\mathbb{R}^2}\left(\frac{1}{n_\infty^2}-\frac{1}{n^2}\right)dx \geqq 0; \tag{4.10}$$

then $\lambda_1(\beta)\leqq\lambda_2(\beta)<(\beta/n_\infty)^2$ *for all* $\beta>0$. *Thus for every nonzero value of* β *there are two eigenvalues of* C_β *below* $(\beta/n_\infty)^2$.

Proof. By [8] the functions $\lambda_m(\beta)$ admit the following equivalent expression:

$$\lambda_m(\beta)=\inf_{V_m\in\mathcal{V}_m}\sup_{\substack{H\in V_m\\ |H|_2=1}} c(\beta; H, H), \tag{4.11}$$

where $\mathcal{V}_m$ is the set of all m-dimensional subspaces of V.

If $v\in H^1(\mathbb{R}^2)$, it results from (4.11) that

$$\lambda_2(\beta)\leqq\sup_{\substack{H\in V_2\\ |H|_2=1}} c(\beta; H, H),$$

where

$$V_2=\left\{\begin{pmatrix}\gamma_1 v\\ \gamma_2 v\\ 0\end{pmatrix};\ (\gamma_1,\gamma_2)\in\mathbb{C}^2\right\}.$$

Since the function v can be chosen arbitrarily, this implies that

$$\lambda_2(\beta)\leqq\left(\frac{\beta}{n_\infty}\right)^2+\mu(\beta), \tag{4.12}$$

where

$$\mu(\beta)=\inf_{\substack{v\in H^1(\mathbb{R}^2)\\ |v|_2=1}}\left\{\frac{1}{n_-^2}\int_{\mathbb{R}^2}|\operatorname{grad} v|^2\,dx-\beta^2\int_{\mathbb{R}^2}\left(\frac{1}{n_\infty^2}-\frac{1}{n^2}\right)|v|^2\,dx\right\}.$$

Suppose first that (4.10) is a strict inequality and consider the following test functions (cf. [19]):

$$v_N(x)=\begin{cases}1 & \text{if } |x|<a,\\ \dfrac{\log|x|-\log N}{\log a-\log N} & \text{if } a<|x|<N,\\ 0 & \text{if } |x|>N,\end{cases}$$

where a is the radius of a disk containing the core region Ω. Then, by (4.12), we have

$$\mu(\beta)\leqq\frac{2\pi\left(n_-^2\log\left(\dfrac{N}{a}\right)\right)^{-1}-\beta^2\displaystyle\int_{D_a}\left(\frac{1}{n_\infty^2}-\frac{1}{n^2}\right)dx}{|v_N|_2^2}.$$

Consequently, by taking N large enough, we see that $\mu(\beta)$ is strictly negative for every nonzero value of β.

If (4.10) is an equality, we introduce some function $w \in H^1(\mathbb{R}^2)$ such that

$$\int_{D_a} \left(\frac{1}{n_\infty^2} - \frac{1}{n^2}\right) w \, dx > 0,$$

and we consider the following test functions:

$$v_N^\alpha = v_N + \alpha w,$$

where α denotes a strictly positive real function. Then, by (4.12), we obtain

$$\mu(\beta) \leqq \frac{4\pi\left(n_-^2 \log\left(\frac{N}{a}\right)\right) + \alpha^2 K(\beta, w) - \alpha\beta^2 \int_{D_a} \left(\frac{1}{n_\infty^2} - \frac{1}{n^2}\right) w \, dx}{|v_N^\alpha|_2^2},$$

where $K(\beta, w)$ neither depends on α or on N. The required result follows by taking N large enough and α small enough. □

Remarks 4.2. (i) A similar result has been established for the scalar equation of the weak-guidance approximation (cf. [6]).

(ii) The converse will be discussed in § 5 (cf. Proposition 5.1).

4.3. Asymptotic behavior of the dispersion curves. Definition of the cutoff values. In this section we prove that, for every integer m, the mth guided mode, associated with the mth eigenvalue $\lambda_m(\beta)$, exists for β large enough. Indeed, we have the following proposition.

PROPOSITION 4.2. *For every integer m there is a value β_m^* such that*

$$\lambda_m(\beta) < (\beta/n_\infty)^2 \quad \forall \beta > \beta_m^* \quad \text{and} \quad \lambda_m(\beta_m^*) = (\beta_m^*/n_\infty)^2.$$

This value is called, by definition, the mth upper cutoff value.

This proposition can be deduced from Lemma 4.2.

LEMMA 4.2. *For every integer m, we have*

$$\lim_{\beta \to +\infty} \frac{\lambda_m(\beta)}{\beta^2} = \frac{1}{n_+^2}. \tag{4.13}$$

Proof. Let $\eta > 0$. By the definition of n_+ there exist $x_0 \in \mathbb{R}^2$ and $\rho > 0$ such that

$$\frac{1}{\rho^2} \int_{D_\rho} \left(\frac{1}{n^2} - \frac{1}{n_+^2}\right) dx < \eta,$$

where $D_\rho = \{x \in \mathbb{R}^2; |x - x_0| < \rho\}$.

We denote by $\mu_\rho^{(1)}, \cdots, \mu_\rho^{(m)}$ the m first eigenvalues of the Laplacian operator in D_ρ with Dirichlet boundary conditions and by $w_\rho^{(m)}$ the associated eigenfunctions extended by zero out of D_ρ and satisfying $|w_\rho^i|_2 = 1$, $i = 1, \cdots, m$.

Let $\tilde{V}_\rho$ be the m-dimensional subspace of $H^1(\mathbb{R}^2)$ spanned by $w_\rho^{(1)}, \cdots, w_\rho^{(m)}$. We set

$$V_\rho = \left\{ \begin{pmatrix} v \\ w \\ 0 \end{pmatrix}; \, v, w \in \tilde{V}_\rho \right\}.$$

By (4.11), we have

$$\lambda_{2m}(\beta) \leqq \sup_{\substack{H \in V_\rho \\ |H|_2 = 1}} c(\beta; H, H),$$

and thus

$$0<\frac{\lambda_{2m}(\beta)}{\beta^2}-\frac{1}{n_+^2}\leqq\alpha_m(\beta,\rho)$$

where

$$\alpha_m(\beta,\rho)=\sup_{\substack{v\in\tilde{V}_\rho\\|v|_2=1}}\left\{\frac{1}{\beta^2 n_-^2}\int_{\mathbb{R}^2}|\operatorname{grad} v|^2\,dx+\int_{\mathbb{R}^2}\left(\frac{1}{n^2}-\frac{1}{n_+^2}\right)|v|^2\,dx\right\}.$$

Finally we derive the following estimate:

$$\left|\frac{\lambda_{2m}(\beta)}{\beta^2}-\frac{1}{n_+^2}\right|\leqq\frac{\mu_\rho^{(m)}}{\beta^2 n_-^2}+\left\{\sup_{\substack{v\in\tilde{V}_\rho\\|v|_2=1}}\|v\|^2_{L^\infty(D_\rho)}\right\}\rho^2\eta.$$

To conclude, we just note that for every integer l

$$\mu_\rho^{(l)}=\frac{1}{\rho^2}\mu_1^{(l)}\quad\text{and}\quad\|w_\rho^{(l)}\|_{L^\infty(D_\rho)}=\frac{1}{\rho}\|w_1^{(l)}\|_{L^\infty(D_1)},$$

so that, for $\beta>1/\rho$

$$\left|\frac{\lambda_{2m-1}(\beta)}{\beta^2}-\frac{1}{n_+^2}\right|\leqq\left|\frac{\lambda_{2m}(\beta)}{\beta^2}-\frac{1}{n_+^2}\right|\leqq K(m)\eta,$$

where $K(m)$ is independent of β, ρ, and η. □

Remarks 4.3. (i) Lemma 4.2 has its own interest. It means that when β tends to infinity the phase velocity of every mode tends to a limit, independent of m and equal to c_0/n_+.

(ii) Other results about the asymptotic behavior of the functions $\lambda_m(\beta)$ have been carried out in [4] and [5].

Proposition 4.2 does not ensure that $\lambda_m(\beta)=(\beta/n_\infty)^2$ when $\beta<\beta_m^*$. That is the reason we set Definition 4.1.

DEFINITION 4.1. For every integer m,

$$\beta_m^0=\sup\left\{\beta_m\in\mathbb{R}^+\;\middle|\;\forall\beta\leqq\beta_m,\ \lambda_m(\beta)=\frac{\beta^2}{n_\infty^2}\right\}.$$

The value β_m^0 is called the mth lower cutoff value.

Obviously, the following inequalities hold:

$$\beta_m^*\leqq\beta_{m+1}^*,\quad\beta_m^0\leqq\beta_{m+1}^0,\quad 0\leqq\beta_m^0\leqq\beta_m^*,$$

and we have (cf. Fig. 3) $\lambda_m(\beta)=\beta^2/n_\infty^2$ for $\beta\leqq\beta_m^0$ and $\lambda_m(\beta)<\beta^2/n_\infty^2$ for $\beta>\beta_m^*$.

These cutoff values will be studied in § 5.

4.4. The case $n\geqq n_\infty$. We can improve these results in the important case where

$$n(x)\geqq n_\infty\quad\text{a.e. in }\mathbb{R}^2. \tag{4.14}$$

Indeed, we have the following theorem.

THEOREM 4.2. *If* (4.14) *holds, then, for every integer m,* $\beta_m^0=\beta_m^*$, *and the number of eigenvalues of* C_β, *located below* $(\beta/n_\infty)^2$, *is monotone nondecreasing in* β.

This theorem, illustrated by Fig. 4, is a straightforward consequence of the next lemma.

LEMMA 4.3. *Assume that* (4.14) *holds. Then the functions* $\Lambda_m(\beta)$, *defined by* (4.3), *are monotone nonincreasing in* β *for* $\beta\in\mathbb{R}^+$.

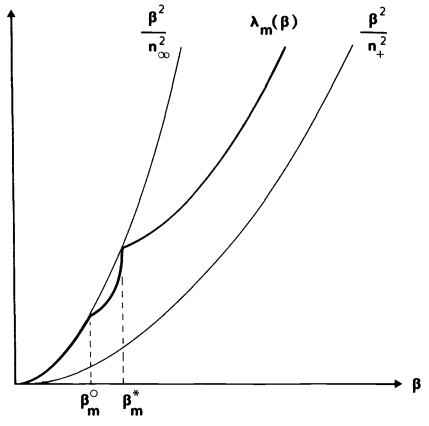


FIG. 3

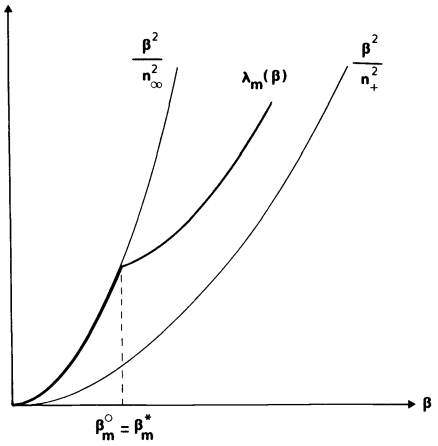


FIG. 4

Proof. We extend to the quadratic case a technique of [22] applied to the linear case.

By (3.12), $\Lambda_m(\beta)$ is always negative, so that the following identity holds:

$$\Lambda_m(\beta) = \sup_{H^{(1)},\cdots,H^{(m-1)} \in (L^2(\mathbb{R}^2))^3} \inf_{\substack{|H|_2=1 \\ H \in [H^{(1)},\cdots,H^{(m-1)}]_V^\perp}} \min(0, d(\beta; H, H)),$$

where $d(\beta; H, H)$ is defined by (3.6). Moreover, for every fixed H, the function which associates $\min(0, d(\beta; H, H))$ to β is monotone nonincreasing in β for $\beta \in \mathbb{R}^+$ (cf. Fig. 5). Indeed, by (4.14), we have

$$d_2(H, H) = \int_{\mathbb{R}^2} \left(\frac{1}{n^2} - \frac{1}{n_\infty^2}\right) |\mathbf{H}|^2 \, dx \leqq 0.$$

The required result follows. □

The structure of the curves $\lambda_m(\beta)$ is, in this case, similar to that of the circular step-index fiber.

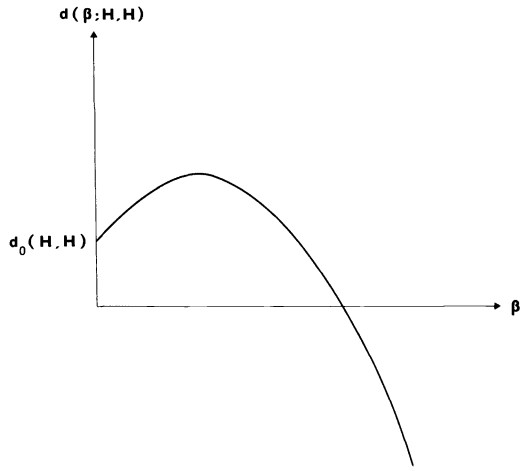


FIG. 5

5. Study of the cutoff equation. The cutoff values, which we introduced in § 4.3, are important features of the guided modes. They are especially useful for counting the number of guided modes which can be propagated in the fiber for a given value of β. Therefore we need a characterization of these cutoff values.

After deriving some a priori estimates, we study the eigenvalue equation

$$c(\beta;\, H_m(\beta), H') = \lambda_m(\beta)(H_m(\beta), H')$$

for β in a neighborhood of β_m^0 or β_m^*. So we obtain the cutoff equation which is set in a weighted Sobolev space.

Finally, we prove that the sequences (β_m^0) and (β_m^*) tend to infinity, and consequently, that the number of guided modes, for a given value of β, is finite.

5.1. A priori estimates. We first establish some a priori estimates for the eigenfield H.

LEMMA 5.1. *Let* $\lambda \in \mathbb{R}^+$ *and* $H \in V$ *such that*

$$c(\beta;\, H, H) = \lambda |H|^2, \qquad \lambda \leqq (\beta / n_\infty)^2. \tag{5.1}$$

Then

$$\int_{\mathbb{R}^2} |\text{grad } H|^2 \, dx \leqq \beta^2 M(n) \int_\Omega |\mathbf{H}|^2 \, dx, \tag{5.2}$$

$$((\beta/n_\infty)^2 - \lambda) \int_{\mathbb{R}^2} |H|^2 \, dx \leqq \beta^2 \Delta(n)(1 + 2M(n)^{1/2}) \int_\Omega |\mathbf{H}|^2 \, dx, \tag{5.3}$$

$$\int_{\mathbb{R}^2} |H_3|^2 \, dx \leqq M(n) \int_\Omega |\mathbf{H}|^2 \, dx, \tag{5.4}$$

where $M(n) = [\Delta(n) n_+^2 + (\Delta(n) n_+^2 + \Delta(n)^2 n_+^4)^{1/2}]^2$, $\Delta(n)$ *is defined by* (4.2), *and* Ω *denotes the core region.*

Proof. By hypothesis, the field H satisfies

$$d_0(H, H) \leqq \beta |d_1(H, H)| + \beta^2 |d_2(H, H)|, \quad \text{and} \tag{5.5}$$

$$((\beta/n_\infty)^2 - \lambda)|H|_2^2 \leqq \beta |d_1(H, H)| + \beta^2 |d_2(H, H)|. \tag{5.6}$$

Moreover, by the Cauchy-Schwarz inequality

$$|d_1(H,H)| \leqq 2\Delta(n)\left(\int_\Omega |\operatorname{grad} H_3|^2\,dx\right)^{1/2}, \quad \text{and} \tag{5.7}$$

$$|d_2(H,H)| \leqq \Delta(n)\int_\Omega |\mathbf{H}|^2\,dx. \tag{5.8}$$

By inserting (5.6), (5.7), and (2.7) in (5.4) and (5.5), we obtain

$$\frac{1}{n_+^2}\int_{\mathbb{R}^2} |\operatorname{grad} H|^2\,dx \leqq \beta^2\Delta(n)\int_\Omega |\mathbf{H}|^2\,dx + 2\beta\Delta(n) \cdot \left(\int_\Omega |\operatorname{grad} H|^2\,dx\right)^{1/2}\left(\int_\Omega |\mathbf{H}|^2\,dx\right)^{1/2} \tag{5.9}$$

and

$$((\beta/n_\infty)^2-\lambda)\int_{\mathbb{R}^2} |H|^2\,dx \leqq \beta^2\Delta(n)\int_\Omega |\mathbf{H}|^2\,dx + 2\beta\Delta(n) \cdot \left(\int_\Omega |\operatorname{grad} H|^2\,dx\right)^{1/2}\left(\int_\Omega |\mathbf{H}|^2\,dx\right)^{1/2}. \tag{5.10}$$

Now we set $X^2 = \int_{\mathbb{R}^2} |\operatorname{grad} H|^2\,dx$ and $Y^2 = \int_\Omega |\mathbf{H}|^2\,dx$. With this notation, (5.7) reads

$$X^2 - 2\beta\Delta(n)n_\infty^2 XY - \beta^2\Delta(n)n_+^2 Y^2 \leqq 0,$$

which implies

$$X \leqq \{\beta\Delta(n)n_+^2 + (\beta^2\Delta(n)^2 n_+^4 + \beta^2\Delta(n)n_+^2)^{1/2}\}\,Y.$$

Inequality (5.2) is exactly the square of the previous inequality. Then we deduce inequality (5.3) from (5.2) and (5.10). Eventually, we obtain inequality (5.4) by using identity (2.2). □

Remarks 5.1. (i) These estimates are satisfied, in particular, if λ is an eigenvalue of the operator C_β and H an associated eigenvector.

(ii) Inequality (5.3) provides a lower bound for the following factor:

$$\frac{\int_\Omega |H|^2\,dx}{\int_{\mathbb{R}^2} |H|^2\,dx},$$

which gives an estimation of the confinement of the field in the core region.

We can deduce from the previous lemma some results concerning the existence of guided modes for small values of the propagation constant β as follows.

PROPOSITION 5.1. (i) *The lower cutoff value of the third guided mode β_0^3 is strictly positive for every index profile.*

(ii) *If the index profile satisfies*

$$\frac{1}{|\Omega|}\int_\Omega \left(\frac{1}{n_\infty^2}-\frac{1}{n^2}\right)dx \geqq -\Delta(n)^2 n_+^2, \tag{5.11}$$

where $|\Omega|$ denotes the measure of Ω, the lower cutoff value of the first mode β_0^1 is strictly positive.

Remarks 5.2. (1) By (i), a fiber supports at most two guided modes at low frequencies.

(2) Assertion (ii) provides a partial converse part for Theorem 4.1.

Proof. (i) Suppose that $\beta_0^3 = 0$ and consider a sequence β_p tending to zero as p tends to infinity and satisfying

$$\lambda_3(\beta_p) < (\beta_p / n_\infty)^2 \tag{5.12}$$

for every p. By (3.11), we have for every value of β

$$\lambda_3(\beta) \geqq \inf_{\substack{H \in \tilde{V} \\ H \neq 0}} \frac{c(\beta; H, H)}{|H|_2^2}, \tag{5.13}$$

where $\tilde{V}$ is the subspace of V defined as follows:

$$\tilde{V} = \left\{ H \in V; \int_\Omega H_1 \, dx = \int_\Omega H_2 \, dx = 0 \right\}.$$

Consequently, by (5.12) and (5.13), there exists a sequence $H^{(p)}$ in $\tilde{V}$ such that

$$c(\beta_p; H^{(p)}, H^{(p)}) < (\beta / n_\infty)^2 |H^{(p)}|_2^2.$$

By Lemma 5.1, sequence $H^{(p)}$ can be normalized as follows:

$$\int_\Omega |\mathbf{H}^{(p)}|^2 \, dx = 1. \tag{5.14}$$

By (5.2), the sequence $\mathbf{H}^{(p)}$ is bounded in $(H^1(\Omega))^2$, and there exists a subsequence converging to $\mathbf{H}$, weakly in $(H^1(\Omega))^2$ and strongly in $(L^2(\Omega))^2$. The limit $\mathbf{H}$ satisfies

$$\int_\Omega |\mathbf{H}|^2 \, dx = 1 \quad \text{and} \quad \int_\Omega H_1 \, dx = \int_\Omega H_2 \, dx = 0. \tag{5.15}$$

Since the sequence β_p tends to zero, it results from (5.2) that H is constant on Ω. The contradiction follows eventually from (5.15).

(ii) Suppose likewise that $\beta_0^1 = 0$ and consider a sequence (β_p), converging to zero and satisfying $\lambda_1(\beta_p) < (\beta_p / n_\infty)^2$ for every p. As in the first part of the proof, we can prove that there exists a sequence (H^p) of V such that

$$c(\beta_p, H^p, H^p) < (\beta_p / n_\infty)^2 |H^p|_2^2 \quad \text{and} \quad \int_\Omega |\mathbf{H}^p|^2 \, dx = 1,$$

for every p, and such that sequence $(\mathbf{H}^{(p)})$ converges strongly to a constant field $\mathbf{H}$ in $(L^2(\Omega))^2$. Moreover, by (5.5) and (5.6), we have

$$\int_{\mathbb{R}} \left(\frac{1}{n_\infty^2} - \frac{1}{n^2} \right) |\mathbf{H}^p|^2 \, dx \geqq \frac{1}{\beta^2 n_+^2} |\operatorname{grad} H^p|_2^2 - \frac{2\Delta(n)}{\beta} |\operatorname{grad} H^p|_2,$$

where $\Delta(n)$ is defined by (4.2) and the term at the right of the inequality is always greater than $(-\Delta(n)^2 n_+^2)$. To conclude, we take the limit of previous inequality when p tends to infinity. □

5.2. The cutoff equation. Assume that $\beta > \beta_m^*$, let $H_m(\beta)$ be an eigenfunction of C_β associated with the eigenvalue $\lambda_m(\beta)$, and consider the eigenvalue equation

$$c(\beta; H_m(\beta), H') = \lambda_m(\beta)(H_m(\beta), H'). \tag{5.16}$$

Formally, when β tends to β_m^*, this equation becomes

$$d(\beta; H_m^*, H') = 0,$$

where we denote by H_m^* the limit, in some sense, of the family $(H_m(\beta))_\beta$. This equation is precisely the so-called cutoff equation.

Now our aim is to make this formal reasoning rigorous by introducing some appropriate functional space.

Note that the limit H_m^* does not generally belong to V since the energy $\int_{\mathbb{R}^2} |\mathbf{H}|^2\, dx$ is generally infinite. Indeed, for $\beta > \beta_m^*$, $H_m(\beta)$ has an exponential decay at infinity (cf. [11]):

$$H_m(\beta)(x) \underset{|x|\to+\infty}{\sim} K\frac{1}{|x|^{1/2}} \exp\{-((\beta/n_\infty)^2 - \lambda_m(\beta))^{1/2}|x|\}$$

but it decreases less and less when $((\beta/n_\infty)^2 - \lambda_m(\beta))$ becomes smaller. At cutoff, it may happen that the field does not decrease outside the core region Ω.

In view of the estimates (5.1) and (5.3), we are led to consider the following functional norm:

$$[H] = \left(\int_\Omega |\mathbf{H}|^2\, dx + \int_{\mathbb{R}^2} |H_3|^2\, dx + \int_{\mathbb{R}^2} |\text{grad } H|^2\, dx\right)^{1/2}. \tag{5.17}$$

We denote by W the completion of $\mathcal{D}(\mathbb{R}^2)$ with respect to this norm. In fact, the space W is equal to the following product space:

$$W = W_0^1(\mathbb{R}^2) \times W_0^1(\mathbb{R}^2) \times H^1(\mathbb{R}^2), \tag{5.18}$$

where $W_0^1(\mathbb{R}^2)$ is a weighted Sobolev space (cf. [15]) defined by

$$W_0^1(\mathbb{R}^2) = \{\varphi;\ \rho\varphi \in L^2(\mathbb{R}^2) \quad \text{and} \quad \text{grad } \varphi \in [L^2(\mathbb{R}^2)]^2\}$$

with $\rho(x) = ((1+|x|^2)^{1/2} \log(2+|x|^2))^{-1}$.

Now we set the following definition.

DEFINITION 5.1. We say that β satisfies the cutoff equation if and only if there exists $\mathrm{H} \in \mathrm{W}$, $\mathrm{H} \neq 0$, such that

$$d(\beta; H, H') = 0 \quad \forall H' \in W,$$

where W is defined by (5.18) and $d(\beta; H, H')$ by (3.6).

This terminology is justified by Theorem 5.1.

THEOREM 5.1. *The cutoff values β_m^0 and β_m^* satisfy the cutoff equation for every integer m.*

Remark 5.3. It is not proved that every solution of the cutoff equation is, conversely, a cutoff value. Nevertheless, this result has been established for the scalar equation (cf. [4]).

Proof. Let $(\beta^p)_{p\in\mathbb{N}}$ be a decreasing sequence tending to β_m^* as p tends to infinity. By the definition of β_m^*, since β^p is greater than β_m^*, $\lambda_m(\beta^p)$ is an eigenvalue of $C_{(\beta^p)}$, and we denote by $H^{(p)}$ an associated eigenfield.

By Lemma 5.1 we can choose the sequence $(H^{(p)})_{p\in\mathbb{N}}$ such that $\int_\Omega |\mathbf{H}^{(p)}|^2\, dx = 1$. Moreover, by (5.1) and (5.3), the sequence $(H^{(p)})$ is bounded in W. Consequently, there exists a subsequence $(H^{(p')})$ and an element H^* in W such that $H^{(p')} \to H^*$ weakly in W and strongly in $(L^2(\Omega))^3$.

We can now take the limit of (5.16) for $H' \in [\mathcal{D}(\mathbb{R}^2)]^3$ and we obtain

$$d(\beta_m^*; H^*, H') = 0,$$

which is valid by density for every H' in W. Finally, H^* does not vanish everywhere since

$$\int_{D_a} |\mathbf{H}^*|^2\, dx = 1.$$

We have proved the lemma for β_m^*. The proof is similar for β_m^0. Indeed, by the definition of β_m^0, there exists a sequence (β^p) such that $\beta^p > \beta_m^0$, $\lambda_m(\beta^p) < (\beta_p/n_\infty)^2$ and $\lim_{p\to+\infty} \beta^p = \beta_m^0$. □

The following lemma is a useful generalization of Lemma 5.1.

LEMMA 5.2. *Let $\beta \in \mathbb{R}^+$ and $H \in W$ such that $d(\beta; H, H) = 0$. Then*

$$\int_{\mathbb{R}^2} |\text{grad } H|^2 \, dx \leqq \beta^2 M(n) \int_\Omega |\mathbf{H}|^2 \, dx,$$

$$\int_{\mathbb{R}^2} |H_3|^2 \, dx \leqq M(n) \int_\Omega |\mathbf{H}|^2 \, dx.$$

We deduce from Lemma 5.2 the following corollary.

COROLLARY 5.3. *For every solution β of the cutoff equation, the associated eigenspace $W_\beta = \{H \in W;\ d(\beta; H, H') = 0,\ \forall H' \in W\}$ is finite-dimensional.*

Remark 5.4. By the previous lemma, if $(\beta/n_\infty)^2$ is an eigenvalue of C_β, it has necessarily finite multiplicity.

Proof. Suppose that W_β is infinite-dimensional and let $(H^{(p)})_p$ be a basis of W_β. By Lemma 5.2, this sequence may be orthogonalized as follows:

$$\int_\Omega \mathbf{H}^{(p)} \cdot \mathbf{H}^{(q)} \, dx = \delta_{pq}. \tag{5.19}$$

Moreover, it is bounded in W and there exists a subsequence of $(\mathbf{H}^{(p)})$ which converges strongly in $(L^2(\Omega))^2$. But this is in contradiction to (5.19). □

5.3. The finite number of eigenvalues. We will now prove that the number of solutions k of the dispersion relation (3.13), for a given value of β, is finite. We first establish the result for a particular category of profiles and then generalize it to arbitrary profiles.

THEOREM 5.2. *Assume that* (4.14) *holds. Then* $\lim_{m\to+\infty} \beta_m^* = +\infty$, *and for each fixed β, C_β has at most a finite number of eigenvalues below $(\beta/n_\infty)^2$.*

Proof. We will prove this result by contradiction. Suppose that sequence (β_m^*) is bounded. Since it is a nondecreasing sequence, there exists β^* such that $\lim_{m\to+\infty} \beta_m^* = \beta^*$.

(1) First we prove that every β greater than β^* satisfies the cutoff equation.

Indeed, if $\beta > \beta^*$, then $\beta > \beta_m^*$ for every integer m. Consequently, the operator C_β has an infinite sequence of eigenvalues $(\lambda_m(\beta))_{m\geqq 1}$ which converges to $(\beta/n_\infty)^2$ as m tends to infinity. Then, using a sequence of associated eigenfields, we can prove, by following the demonstration of Theorem 5.1, that β satisfies the cutoff equation.

(2) Let $(\beta^p)_{p\in\mathbb{N}}$ be a strictly decreasing sequence tending to β^* as p tends to infinity. By the previous paragraph of the proof, β^p satisfies the cutoff equation for every p. Let us denote by $H^{(p)}$ an associated eigenfield which is assumed to satisfy (cf. Lemma 5.2)

$$\int_\Omega |\mathbf{H}^{(p)}|^2 \, dx = 1.$$

By Lemma 5.1, there exists a subsequence, still denoted $(H^{(p)})_{p\in\mathbb{N}}$, and an element H^* in W such that $H^{(p)} \to H^*$ weakly in W and strongly in $(L^2(\Omega))^3$. The field H^* satisfies

$$\int_\Omega |\mathbf{H}^*|^2 \, dx = 1 \quad \text{and} \quad d(\beta^*; H^*, H') = 0 \quad \forall H' \in W. \tag{5.20}$$

(3) Let $p, q \in \mathbb{N}$. By definition of $H^{(p)}$, we have $d(\beta^p; H^{(p)}, H^{(q)}) = 0$ and $d(\beta^q; H^{(q)}, H^{(p)}) = 0$.

By subtracting the conjugate part of the second equation from the first one and by dividing the result by $(\beta_p - \beta_q)$, we get $d_1(H^{(p)}, H^{(q)}) + (\beta^p + \beta^q) d_2(H^{(p)}, H^{(q)}) = 0$, which converges to

$$d_1(H^*, H^*) + 2\beta^* d_2(H^*, H^*) = 0, \tag{5.21}$$

as p tends to infinity. By (5.20) we have, in addition,

$$d_0(H^*, H^*) + \beta^* d_1(H^*, H^*) + (\beta^*)^2 d_2(H^*, H^*) = 0. \tag{5.22}$$

Using (5.20)-(5.22), we eventually obtain

$$d_0(H^*, H^*) = (\beta^*)^2 d_2(H^*, H^*).$$

But we have $d_0(H^*, H^*) \geqq 0$ and under assumption (4.14), $d_2(H^*, H^*) < 0$. Eventually, by Proposition 5.1, $\beta^* > 0$. □

Using a comparison technique, we can generalize this result as follows.

THEOREM 5.3. *The sequences of cutoff values* (β_m^0) *and* (β_m^*) *satisfy*

$$\lim_{m \to +\infty} \beta_m^0 = +\infty, \qquad \lim_{m \to +\infty} \beta_m^* = +\infty$$

and for each fixed β, C_β *has at most a finite number of eigenvalues below* $(\beta / n_\infty)^2$.

Proof. Assume that the index profile n does not satisfy (4.14) and let us define the index profile $\bar{n}$ of a step-index fiber by

$$\bar{n}(x) = n_+ \quad \text{if } x \in \Omega,$$
$$\bar{n}(x) = n_\infty \quad \text{if } x \in \mathbb{R}^2 \backslash \Omega.$$

Since $\bar{n}(x) \geqq n(x)$, almost everywhere $x \in \mathbb{R}^2$, by Remark 3.3, $\lambda_m(\beta, \bar{n}) \leqq \lambda_m(\beta, n)$ for every integer m and every value of β. Consequently, by definition of the cutoff values:

$$\beta_m^0(\bar{n}) \leqq \beta_m^*(\bar{n}) \leqq \beta_m^0(n) \leqq \beta_m^*(n).$$

Moreover, $\bar{n}$ satisfies (4.14). Therefore, by Proposition 4.2 $\beta_m^0(\bar{n}) = \beta_m^*(\bar{n})$, and by Theorem 5.2, $\lim_{m \to +\infty} \beta_m^*(\bar{n}) = +\infty$. □

6. Conclusion. This work provides a relatively complete description of the dispersion relation of a fiber, whose index profile is arbitrary.

The results are more precise when the index is everywhere in the core greater than the index of the cladding. In that case, we showed that the structure of the guided modes is similar to that of a circular step-index fiber.

When the above condition is not satisfied, some points must still be investigated, especially the existence of index profiles such that $\beta_m^0 < \beta_m^*$ for some m, and, in this case, the behavior of $\lambda_m(\beta)$ in the interval $[\beta_m^0, \beta_m^*]$. The cutoff equation therefore requires specific study.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1964.

[2] A. BAMBERGER AND A. S. BONNET, *Calcul des modes guidés d'une fibre optique. Deuxième partie: analyse mathématique*, Rapport interne 143, Centre de Mathématiques Appliquées, Ecole Polytechnique, Palaiseau, France, 1986.

[3] A. BAMBERGER, A. S. BONNET, AND R. DJELLOULI, *Calcul des modes guidés d'une fibre optique. Première partie: différentes formulations mathématiques du problème*, Rapport interne 142, Centre de Mathématiques Appliquées, Ecole Polytechnique, Palaiseau, France, 1986.

[4] A. S. BONNET, *Analyse mathématique de la propagation de modes guidés dans les fibres optiques*, Thèse d'analyse numérique de l'Université Pierre et Marie Curie, Paris, France, 1988.

[5] A. S. BONNET AND R. DJELLOULI, *Etude mathématique des modes guidés d'une fibre optique. Résultats complémentaires et extension au cas de couplages*, Rapport interne 182, Centre de Mathématiques Appliquées, Ecole Polytechnique, Palaiseau, France, 1988.

[6] G. COPPA AND P. DI VITTA, *Cut-off condition of the fundamental mode in monotone fibers*, Optics Comm., 49 (1984), pp. 409–412.

[7] R. DAUTRAY AND J. L. LIONS, *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Tome 2, Masson, Paris, 1985.

[8] N. DUNFORD AND J. SCHWARTZ, *Linear Operators. Part* II: *Spectral Theory*, Interscience, New York, 1963.

[9] J. C. GUILLOT, *Complétude des modes T.E. et T.M. pour un guide d'ondes optiques planaire*, Rapport de Recherche INRIA, Le Chesnay, France, 385, 1985.

[10] L. HORMANDER, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1976.

[11] L. B. JEUNHOMME, *Single-Mode Fiber Optics: Principles and Applications*, Marcel Dekker, New York, 1983.

[12] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, 1976.

[13] M. KOSHIBA, K. HAYATA, AND M. SUZUKI, *Vectorial finite-element method without spurious solutions for dielectric waveguide problems*, Electron. Lett., 20 (1984), pp. 402–410.

[14] R. LEIS, *Zur Theorie Elektromagnetischer Schwingungen in Anisotropen Medien*, Math. Z., 106 (1968), pp. 213–224.

[15] M. N. LE ROUX, *Thèse de 3ème cycle*, Université de Rennes, Rennes, France, 1974.

[16] D. MARCUSE, *Theory of Dielectric Optical Waveguide*, Academic Press, New York, 1974.

[17] ———, *Light Transmission Optics*, Van Nostrand, New York, 1982.

[18] N. MABAYA, P. E. LAGASSE, AND P. VANDENBULCKE, *Finite element analysis of optical waveguides*, IEEE Trans. Microwave Theory Tech., (1981), pp. 600–605.

[19] H. PICQ, *Détermination et calcul numérique de la première valeur propre d'opérateurs de Schrödinger dans le plan*, Thèse, Université de Nice, Nice, France, 1982.

[20] J. P. POCHOLLE, *Propagation dans les fibres optiques monomodes*, Revue technique de Thomson-CSF, 15, 1983.

[21] F. RELLICH, *Über das asymptotische Verhalten der Lösungen von* $\Delta u + \lambda u = 0$ *in unendlichen Gebieten*, Über. Deutsch. Math. Verein, 53 (1943), pp. 57–65.

[22] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics*, Vol. 4, Academic Press, New York, 1978.

[23] M. SCHECHTER, *Operator Methods in Quantum Mechanics*, North Holland, Amsterdam, 1981.

[24] A. W. SNYDER AND J. D. LOVE, *Optical Waveguide Theory*, Chapman and Hall, London, 1983.

[25] C. VASSALO, *Théorie des guides d'ondes électromagnétiques*, Tomes 1 et 2, Editions Eyrolles et CNET-ENST, Paris, 1985.

[26] R. WEDER, *Spectral and scattering theory in deformed optical wave guides*, J. Reine Angew. Math., 390 (1988), pp. 130–169.

[27] C. WILCOX, *Sound Propagation in Stratified Fluids*, Applied Mathematical Sciences, Vol. 50, Springer-Verlag, Berlin, New York, 1984.

[28] C. YEH, K. HA, S. B. DONG, AND W. P. BROWN, *Single-mode optical waveguides*, Appl. Optics, 18 (1979), pp. 1490–1504.

[29] H. ZIANI AND C. DEVYS, *Méthode intégrale pour le calcul des modes guidés d'une fibre optique*, Rapport Interne, Centre de Mathématiques Appliquées, 129, Ecole Polytechnique, Palaiseau, France, 1985.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1511-1536, November 1990
© 1990 Society for Industrial and Applied Mathematics
008

CORRELATIONS BETWEEN CHAOS IN A PERTURBED SINE-GORDON EQUATION AND A TRUNCATED MODEL SYSTEM*

ALAN R. BISHOP†, RANDY FLESCH‡, M. GREGORY FOREST§[1],
DAVID W. MCLAUGHLIN¶, AND EDWARD A. OVERMAN, II§[2]

Abstract. The purpose of this paper is to present a first step toward providing coordinates and associated dynamics for low-dimensional attractors in nearly integrable partial differential equations (pdes), in particular, where the truncated system reflects salient geometric properties of the pde. This is achieved by correlating:

(i) Numerical results on the bifurcations to temporal chaos with spatial coherence of the damped, periodically forced sine-Gordon equation with periodic boundary conditions;

(ii) An interpretation of the spatial and temporal bifurcation structures of this perturbed integrable system with regard to the exact structure of the sine-Gordon phase space;

(iii) A model dynamical systems problem, which is itself a perturbed integrable Hamiltonian system, derived from the perturbed sine-Gordon equation by a finite mode Fourier truncation in the nonlinear Schrödinger limit; and

(iv) The bifurcations to chaos in the truncated phase space.

In particular, a potential source of chaos in both the pde and the model ordinary differential equation systems is focused on: the existence of homoclinic orbits in the unperturbed integrable phase space and their continuation in the perturbed problem. The evidence presented here supports our thesis that the chaotic attractors of the weakly perturbed periodic sine-Gordon system consists of low-dimensional metastable attracting states together with intermediate states that are $O(1)$ unstable and correspond to homoclinic states in the integrable phase space. It is surmised that the chaotic dynamics on these attractors is due to the perturbation of these homoclinic integrable configurations.

Key words. chaos, sine-Gordon equation, homoclinic orbits

AMS(MOS) subject classifications. 35B32, 35B35, 58G20

Introduction. The purpose of this paper is to present a first step toward providing coordinates and associated dynamics for low-dimensional attractors in nearly integrable partial differential equations (pdes). In this paper we describe

(i) Numerical results on the bifurcations of the damped, periodically forced sine-Gordon equation with periodic boundary conditions which reveal attractors that are spatially coherent while temporally chaotic;

(ii) An interpretation of the spatial and temporal bifurcation structures of this perturbed integrable system with regard to the integrable structure of the sine-Gordon phase space;

(iii) A model dynamical systems problem, which is itself a perturbed integrable Hamiltonian system, derived from the perturbed sine-Gordon equation by a finite mode truncation in the nonlinear Schrödinger limit; and

(iv) The bifurcations to chaos in the four-dimensional truncated phase space.

* Received by the editors June 16, 1988; accepted for publication (in revised form) December 1, 1989.

† Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratories, Los Alamos, New Mexico 87545.

‡ L.A.M.F. Technical University of Denmark, Building 303, DK-2800, Lyngby, Denmark.

§ Department of Mathematics, Ohio State University, Columbus, Ohio 43210.

¶ Department of Mathematics and Program in Applied Mathematics, University of Arizona, Tucson, Arizona 85721. The work of this author was partially supported by National Science Foundation grant DMS 8403187 and Air Force Office of Scientific Research grant AFOSR 830227.

[1] The work of this author was partially supported by National Science Foundation grant DMS 8803465.

[2] The work of this author was partially supported by National Science Foundation grant DMS 8818640 and Air Force Office of Scientific Research grant AFOSR 88-0195. This author acknowledges a grant of computer time from the Ohio Supercomputer Center.

In particular, we focus on a likely source of chaos in both the pde and ordinary differential equation (ode) systems: the existence of homoclinic orbits in the unperturbed integrable phase space and their continuation in the perturbed problem. In the last part of this study, we numerically correlate the homoclinic crossings in the chaotic dynamics of the full and reduced problems.

While the present paper does succeed in revealing homoclinic structure of the pde in a finite mode truncation, we do not claim that this four-dimensional real truncation is sufficient for other important features. On the contrary, two more dimensions are required to accurately cover the attractor [11], to resolve the unstable manifolds of metastable states on the attractors [12], and to quantitatively reproduce the pde bifurcation sequence [11].

These and similar [1], [2] experimental results provide information about

(i) The coexistence of simple coherent spatial structures and temporal chaos;

(ii) The potential for capturing the pde bifurcation sequence with truncated modal systems; and

(iii) The potential for identifying coordinates for chaotic attractors.

These studies also provide directions for the rigorous mathematical analysis to support the numerical work in the individual pde and ode systems, as well as to develop the connections between the full and reduced systems. We discuss some current projects in § 6.

The outline for the remainder of the paper is as follows:

Section 1 gives numerical bifurcations of the perturbed sine-Gordon equation; § 2 gives a truncated two-mode expansion in the nonlinear Schrödinger limit; § 3 gives properties of the unperturbed modal equations; § 4 gives bifurcations of the perturbed modal equations; and § 5 gives correlations between the infinite-dimensional and reduced systems.

1. Numerical bifurcations of the perturbed sine-Gordon equation. We begin by describing one particular experiment from our body of numerical studies (e.g., [1], [2]) on the weakly damped, periodically forced, sine-Gordon equation

$$u_{tt} - u_{xx} + \sin u = \varepsilon[-\alpha u_t + \Gamma \cos(\omega t)], \tag{1.1a}$$

under periodic boundary conditions

$$u\left(x = -\frac{L}{2}, t\right) = u\left(x = \frac{L}{2}, t\right) \quad \text{for all } t, \tag{1.1b}$$

and with even spatial symmetry

$$u(x, t) = u(-x, t) \quad \text{for all } t.$$

For the purpose of this paper we restrict attention to one bifurcation parameter $\varepsilon\Gamma$, the amplitude of the external driver. The remaining parameters are fixed in the following way:

(i) The linear damping coefficient $\varepsilon\alpha$ is chosen very small:

$$\varepsilon\alpha = .04; \tag{1.1c}$$

(ii) The external driving frequency ω is chosen near but less than 1:

$$\omega = 1 - \varepsilon\tilde{\omega} = .87; \tag{1.1d}$$

(iii) The spatial period L is fixed at

$$L = 12; \tag{1.1e}$$

and

(iv) The initial condition is given as a single-hump sine-Gordon breather localized within the period.

With this parameter specification, we observe the following (Fig. 1) long-time asymptotic states as a function of the bifurcation parameter $\varepsilon\Gamma$. The numerical methods used to discretize this pde are discussed in the Appendix. These long-time states, or "attractors," are specified by their spatial structure and temporal behavior, with the notation: K_0 denotes a spatially homogeneous component, of zero wavenumber; K_1 denotes a period one component of wavenumber $K_1 = 2\pi/L$; $K_0 \oplus K_1$ denotes the nonlinear superposition of the two modes, etc. *Locked* implies a frequency-locked state, oscillating at the driven frequency ω. Chaotic denotes a broad-banded frequency spectrum.

This particular bifurcation sequence does not exhibit quasi periodicity prior to chaos, which is a typical route to chaos in other parameter regimes [1], [2]. A more exhaustive parameter study is required to resolve whether stable quasi-periodic attractors occur in this diagram. However, the model problem we present below indicates that when a second frequency is excited at this parameter specification, the collective quasi-periodic state is unstable and thus would not be observed numerically. This structure is reflected in the pde chaotic dynamics in that the system intermittently settles into weakly unstable quasi-periodic states; we illustrate this in Fig. 2, where $\varepsilon\Gamma = .103$.

We emphasize the spatial coherence coexisting with temporal intermittent chaos, and moreover in a parameter range very near the integrable sine-Gordon pde. Fig. 2(a) displays the evolution of the spatial structure in time, beginning at $t = 50{,}000$, long after all transients have passed. Note the intermittent jumping between two weakly unstable spatial structures, a "breather" (localized hump) peaked either at the center or at the ends of the interval, with an intermediate passage through a flat state.

In order to quantify this spatial structure at each timestep t_n, we use a recently developed sine-Gordon spectral code to measure the exact sine-Gordon nonlinear mode content in the field $u^\varepsilon(x, t_n)$. (See [2] for details.) For example, Fig. 3(a) is the sine-Gordon spectrum for an exact K_0 sine-Gordon solution, i.e., a solution of the pendulum: $u = 2\sin^{-1}[ksn(t; k)]$, $0 < k^2 < 1$, with frequency $\omega = .87$, and for a spatial period $L = 12$. These spectral curves are invariant under the exact sine-Gordon flow. The endpoints of curves of spectrum are simple periodic spectra, and are closely related to the action variables in the action-angle linearization of periodic sine-Gordon [6]. The other marked points, denoted by $\triangle$ or $\square$ *within* bands of spectrum, are double periodic spectra, and these label all closed (degenerate) degrees of freedom. In [3] we

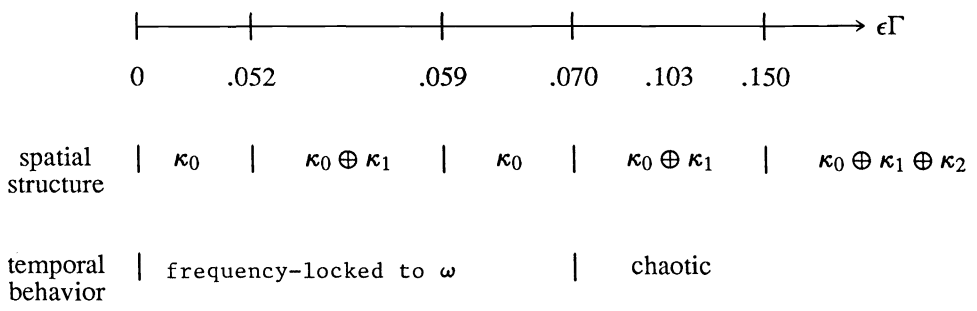


FIG. 1. *The pde bifurcation diagram, corresponding to variable $\varepsilon\Gamma$ with all remaining parameters fixed in equations* (1.1).

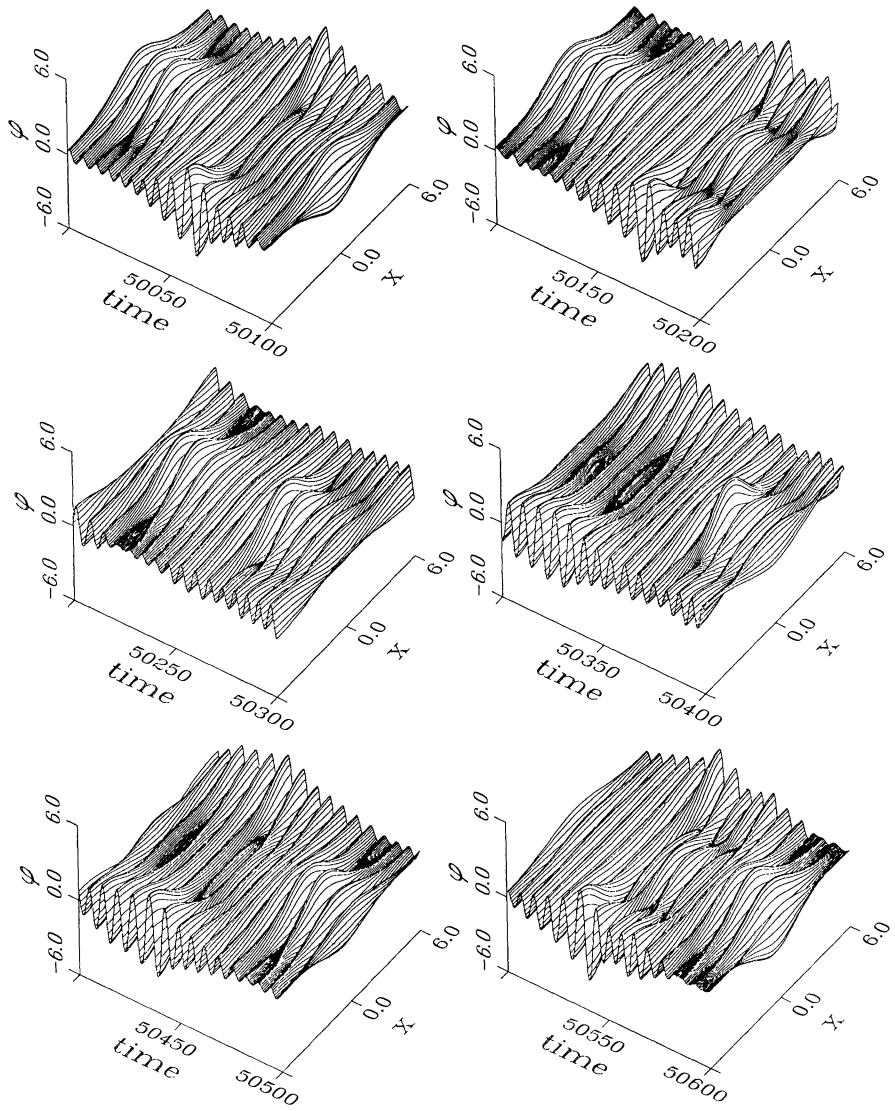


FIG. 2(a). *Numerically computed solution* $\phi = u^{\varepsilon}(x, t)$ *of equation* (1.1), $50{,}000 \leqq t \leqq 50{,}600$.

show that there is a 1:1 correspondence between pairs of fully complex (nonreal) double points and linearized, exponentially unstable modes for a given sine-Gordon N degree of freedom solution. Moreover, these local instabilities are reflected globally in the isospectral set of the given solution by homoclinic components. For the example, Fig. 3(a), the number of pairs of complex double points is given by the integer solutions n of $(2n\pi/L)^2 \in (0, k^2)$, where k is the elliptic modulus. Since $L = 12$ here, we find that $n = 1$ is the only solution. The exact K_0 sine-Gordon solution, depicted in Fig. 3(a), with frequency as in (1.1d), on the interval of length $L = 12$,

(i) Is linearly unstable, with order 1 growth rate;

(ii) Has homoclinic orbits on its sine-Gordon isospectral set, which are homoclinic as $t \to \pm\infty$ to this circle (one-torus) of constants in the phase space of L-periodic functions of x; and

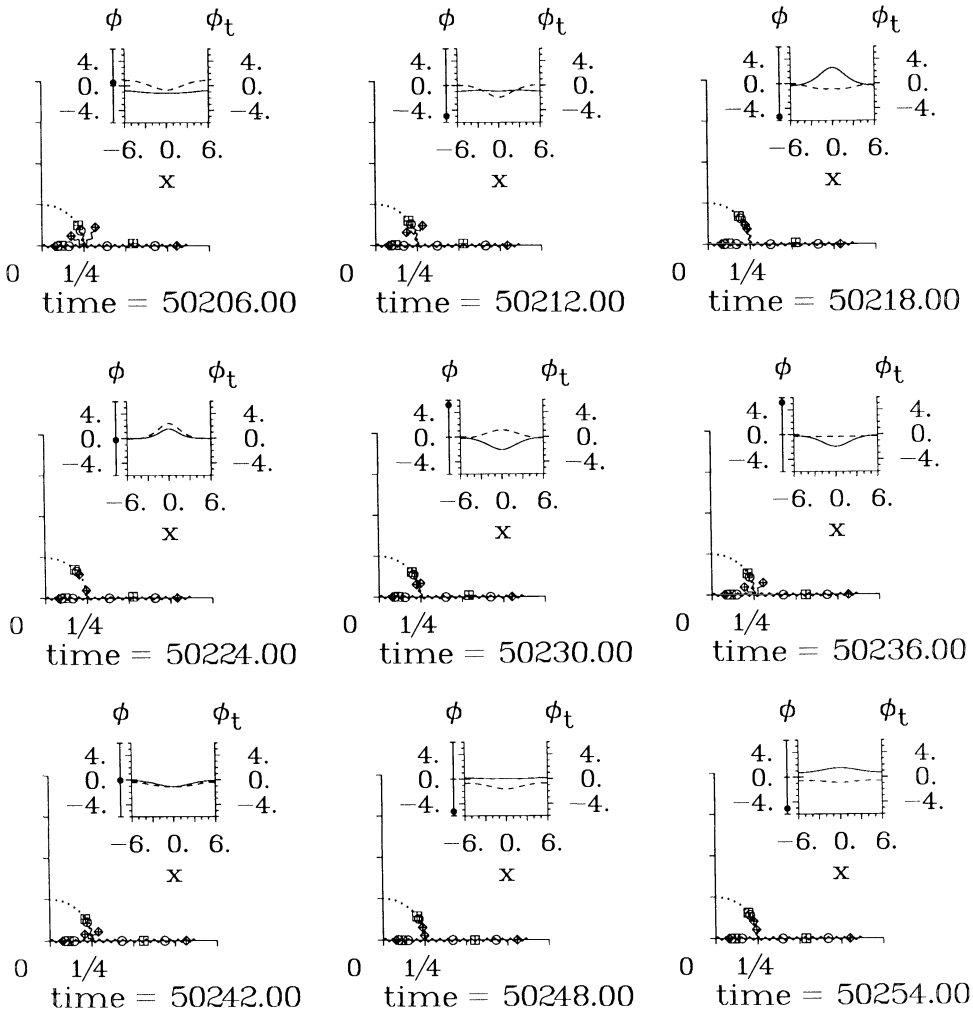
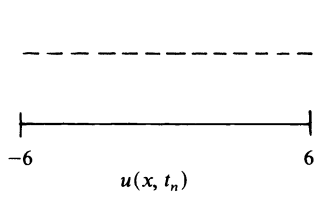


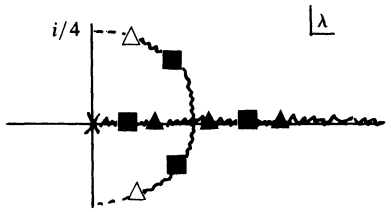
FIG. 2(b). *Corresponding to the numerically computed solution in Fig.* 2(a), $\phi = u^{\varepsilon}(x, t_n)$ *(solid graph) and* $\phi_t = (\partial/\partial t)u^{\varepsilon}(x, t_n)$ *(dotted graph), the sine-Gordon mode projection is numerically measured, with respect to the complex spectral parameter* λ*, at selected times* t_n*. (Refer ahead to Fig.* 3 *for the* λ*-spectral measurement of three exact sine-Gordon low-mode solutions.) We note:* (1) *the passage of the perturbed flow from a "cross" spectral projection to a "gap" spectral measurement; and* (2) *at each discrete time* t_n*, the spatial waveforms are predominantly nonlinear* $K_0 \oplus K_1$ *sine-Gordon waveforms.*

(iii) This instability saturates nonlinearly, by arbitrary variation of initial conditions, to the breathers in our following examples [3].

We emphasize that these are low-amplitude spatial structures, far from the amplitude π associated with inverted rest states of the pendulum. These homoclinic orbits of the full pde are thus quite distinct from the separatrices in the x-independent pendulum equation.

The next example is an exact $K_0 \oplus K_1$ sine-Gordon solution: a breather plus nonzero mean. These are *two* such exact nonlinear states (Fig. 3(b)) reflecting the two ways that the degeneracy due to the complex double point in Fig. 3(a) can break. These nonlinear $K_0 \oplus K_1$ states represent exact sine-Gordon solutions, with frequency

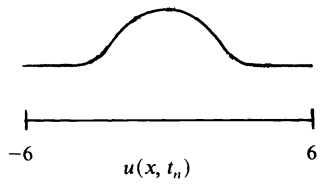


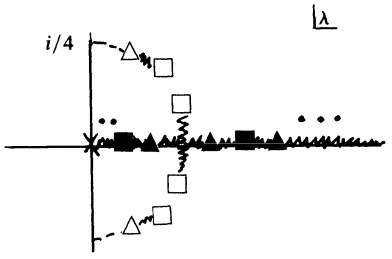


K_0 plus complex pair of double points

denotes a band of spectrum
△, □ denote simple periodic spectra
▲, ■ denote double periodic spectra

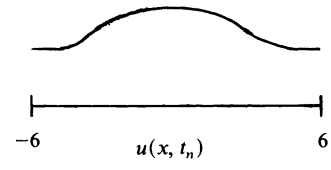
(a)

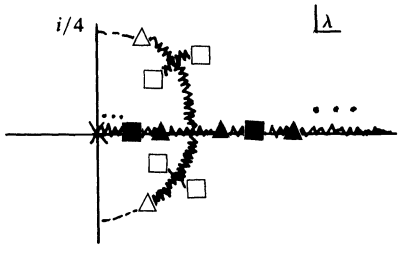




$K_0 \oplus K_1$ "gap" state

(b)





$K_0 \oplus K_1$ "cross" state

(c)

FIG. 3. *A schematic of the spatial structure for three exact sine-Gordon solutions at some discrete time* t_n, $u(x, t_n)$, *together with the associated* λ*-spectral measurement. We remark that for exact sine-Gordon flows, the* λ*-spectral projection is invariant. Fig.* 3(a) *depicts the spectrum of an x-independent pendulum solution which is purely oscillatory in t, with frequency* $\omega = .87$, *on an x-interval of length* $L = 12$. *This pure* K_0 *state has one spectral band on the circle of radius* $\frac{1}{4}$, *emanating from* $\lambda = \frac{1}{4}$, *terminating at simple periodic eigenvalues* $\lambda = e^{\pm i\psi}/16$, *where* ψ *measures the maximum amplitude of u. For this* K_0 *solution, on an interval of length* $L = 12$, *there is one pair of complex conjugate double points which labels the modulational instability of this x-independent solution in the* K_1 *mode direction. Figures* 3(b) *and* 3(c) *depict pure* $K_0 \oplus K_1$ *spatial waveforms of sine-Gordon, so-called "breather plus nonzero mean" states. The two spectral configurations represent the two ways the complex double points of Fig.* 3(a) *may break into simple periodic points, opening up order* 1 *amplitude in the* K_1 *mode, and producing the two types of exact nonlinear* $K_0 \oplus K_1$ *waveforms.*

$\omega = .87$ as in (1.1d), x-period $L = 12$, and are linearly neutrally stable solutions of the sine-Gordon equation (there are no fully complex double points). Under the unperturbed sine-Gordon dynamics, the spectral configurations in Figs. 3(a) and 3(b) remain invariant.

Under a weakly perturbed flow such as (1.1), initial configurations such as in Fig. 3 will distort due to the perturbation. The endpoints of the spectral curves will modulate, and moreover, the "closed" degrees of freedom will be opened. On short timescales, the neutrally stable modes (associated with real double points) perturb only to the order of the perturbation, whereas the $O(1)$ unstable modes (associated to nonreal double points) generate dramatic changes in spatial structure. Of interest here is which modes resonate with significant amplitude on very long timescales, after all transients have passed, and then how the dynamics of these modes proceed.

At each timestep in the perturbed flow (1.1), we measure the exact nonlinear content in $u^\varepsilon(x, t_n)$, $t_n \gg 1$. In this way we determine if the spatial structure is well approximated:

(1) At a given instant $t_n \gg 1$ by a low degree of freedom exact sine-Gordon field, and

(2) During the flow by a slow modulation through the low-dimensional nonlinear modes, or if the mode content varies widely in a sine-Gordon projection.

Figure 2(b), corresponding to Fig. 2(a), indicates the sine-Gordon mode projection of $u^\varepsilon(x, t_n)$, $t_n \gg 1$, is uniformly very low-dimensional, even in this chaotic regime, and that the energy transfer is predominantly within the nonlinear K_0 and K_1 modes. These measurements quantify our spatial description of the bifurcation diagram in Fig. 1.

Remark. In the Fourier mode projection of $u^\varepsilon(x, t_n)$, a second harmonic $\cos(K_2 x)$ is required to accurately describe the weak instabilities of the (metastable) spatial structures that comprise the chaotic attractor. We refer to [12] for this analysis, and to [11] for a discussion of the truncated Fourier mode system that includes this second harmonic.

The next figure, Fig. 4, is a phase-plane projection of (u, u_t) at one location, $x = 0$, again for $\varepsilon\Gamma = .103$. Another indication of the chaotic dynamics is a broadband power spectrum, which we omit here. (Refer to [1] and [2] for an exhaustive description of the dynamical systems tools which we use to measure the frequency locked, quasi-periodic, and chaotic attractors.)

We close this section with a summary description of the chaotic attractors for (1.1). The dynamics settles into a region of phase space containing two nonlinear $K_0 \oplus K_1$ states (breather plus nonzero mean), with the breather localized in the center of the interval, the other state with the breather translated by $L/2$ to the wings. Note the discrete symmetry due to periodic boundary conditions and symmetric initial data. Each state is "unstable," with weak $O(\varepsilon)$ instabilities due to the perturbation, and the unstable flow out of each state is through a neighborhood of the flat K_0 state, landing either back into the original $K_0 \oplus K_1$ mode, or into the translated $K_0 \oplus K_1$ state. The intermediate K_0 state, however, is unstable (with order 1 growth rate) even in the unperturbed flow, where it has homoclinic orbits associated to it. This apparent random jumping process between the two $K_0 \oplus K_1$ states begs to be identified as a Bernoulli shift on two symbols. In this description the two symbols are identified with the neighborhoods of the two $K_0 \oplus K_1$ states, whereas the perturbed homoclinic structure is responsible for the Bernoulli shift on these symbols.

This phenomenon is the sine-Gordon low-amplitude analogue of the now classical larger amplitude pendulum chaos: the exponentially unstable inverted state ($u = \pi$) which under perturbation has equal likelihood of falling into either of the two states

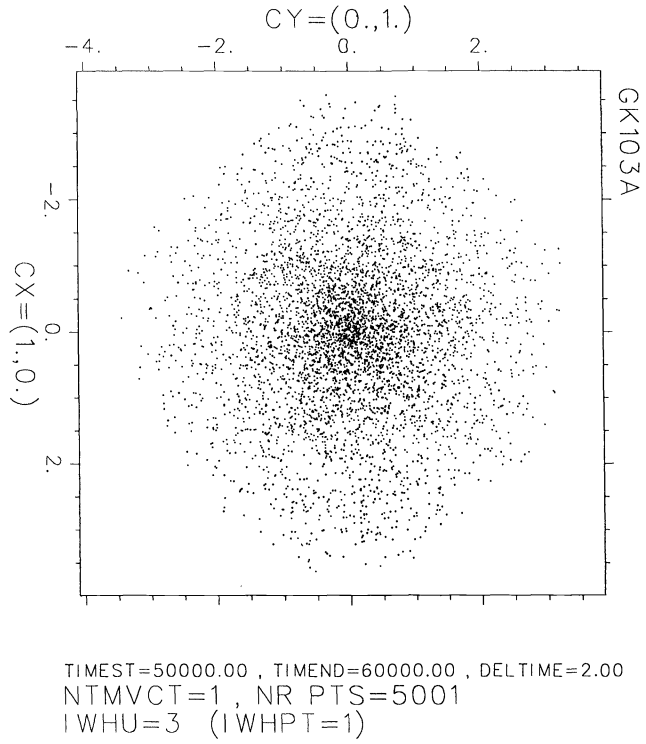


FIG. 4. *The phase plane projection of* (u, u_t) *at* $x=0$, *for the run in Fig.* 2(a), $(u^\varepsilon(0, t_n), u_t^\varepsilon(0, t_n))$, $t_n \in [50{,}000, 50{,}600]$, *with* $\Delta t = 2$.

(oscillatory or running) it separates. These two states are stable in the exact pendulum dynamics, but develop $O(\varepsilon)$ instabilities due to the perturbation, and their perturbed flow often passes randomly through the homoclinic configuration.

Our thesis is that the chaotic attractors of the weakly perturbed periodic sine-Gordon system, e.g., (1.1), consist of low-dimensional metastable attracting states, e.g., the nonlinear $K_0 \oplus K_1$ states of Fig. 2, together with intermediate states that are $O(1)$ unstable and represent homoclinic configurations in the integrable phase space. The chaotic dynamics on these attractors is, we surmise, due to these perturbed homoclinic configurations.

The derivation and analysis of perturbed, fully nonlinear, action-angle modes, truncated on the low-dimensional structures associated to Figs. 2 and 3, are currently under way [5]. As a preliminary step, we develop here a simple model problem that captures some essential qualitative features of this route to chaos in the nearly integrable pde. This model problem is achieved through a natural finite mode truncation on the first two Fourier mode complex amplitudes.

We refer to [11] for a higher mode truncation which is aimed at more accurately covering the chaotic attractors.

2. A truncated finite mode expansion in the nonlinear Schrödinger limit. For frequencies near but less than 1, the weakly perturbed sine-Gordon flow (1.1) resonates with low-amplitude "breatherlike" spatial modes, rather than kink-like modes which can predominate for significantly lower ω [10]. In this limit we easily derive a perturbed nonlinear Schrödinger envelope equation as follows. Seek a solution of (1.1) in the following form (recall $\omega = .87 = 1 - \varepsilon\tilde{\omega}$, (1.1d)):

$$u^\varepsilon = 2\sqrt{\varepsilon\tilde{\omega}}\,[B(X, T)\, e^{i\omega t} + \text{complex conjugate}] + O(\varepsilon), \tag{2.1a}$$

with

$$X = \sqrt{2\varepsilon\tilde{\omega}}\, x, \qquad T = \varepsilon\tilde{\omega} t. \tag{2.1b}$$

Then the slowly varying envelope $B(X, T)$ satisfies

$$-iB_T + B_{XX} + (|B|^2 - 1)B = i\tilde{\alpha}B + \tilde{\Gamma}. \tag{2.2}$$

From (1.1) the scaled parameters become (approximately):

$$\tilde{\alpha} = \frac{\alpha}{2\varepsilon\tilde{\omega}} \approx .154, \quad L_X = 12\sqrt{2\varepsilon\tilde{\omega}} \approx 6.12, \quad k_X = \frac{2\pi}{L_X} \approx 1.025, \tag{2.3a}$$

and the bifurcation parameter $\tilde{\Gamma}$ is now

$$\tilde{\Gamma} = \frac{\varepsilon\Gamma}{8(\varepsilon\tilde{\omega})^{3/2}} \approx 2.67\varepsilon\Gamma. \tag{2.3b}$$

We have achieved two things by reducing to this amplitude equation. First, we preserve the perturbed integrable structure, since the unperturbed pde (2.2) with $\tilde{\alpha} = \tilde{\Gamma} = 0$ is the integrable nonlinear Schrödinger equation. Second, we factor out one frequency, ω of the driver. Thus, steady solutions of (2.2) correspond to frequency locked solutions of (1.12), while T-periodic flows of (2.2), incommensurate with ω, correspond to quasi-periodic perturbed sine-Gordon solutions. Chaos in one system is chaos in the other.

We now make a further approximation and truncation based on the predominant $K_0 \oplus K_1$ structure measured in Figs. 2 and 3 for the perturbed sine-Gordon flow. (A similar truncation and an interesting numerical study appears in [9]. The primary difference is our focus here on the role of homoclinic structures in the attractors and the comparison with the perturbed pde.) We seek

$$B(X, T) = c(T) + b(T)\cos(kX), \qquad k = \frac{2\pi}{L_X}. \tag{2.4}$$

Inserting this ansatz into the perturbed NLS equation (2.2) and retaining cubic terms in the complex Fourier amplitudes $c(T)$, $b(T)$ yields

$$\begin{aligned} -ic_T + (|c|^2 + \tfrac{1}{2}|b|^2 - 1)c + \tfrac{1}{2}(cb^* + c^*b)b &= i\tilde{\alpha}c + i\tilde{\Gamma}, \\ -ib_T + (|c|^2 + \tfrac{3}{4}|b|^2 - (1 + k^2))b + (cb^* + bc^*)c &= i\tilde{\alpha}b. \end{aligned} \tag{2.5}$$

Several remarks about this model four-dimensional dynamical system are appropriate at this point.

Remark 1. This two-complex Fourier mode truncation is surely not expected to yield quantitative agreement with the perturbed pde, although Fig. 3 suggests the two mode $K_0 \oplus K_1$ nonlinear truncations provide a very good approximation [5]. This discrepancy in the linear versus nonlinear mode is apparent as we compare a sine-Gordon breather K_1 mode with the Fourier K_1 mode (Fig. 5).

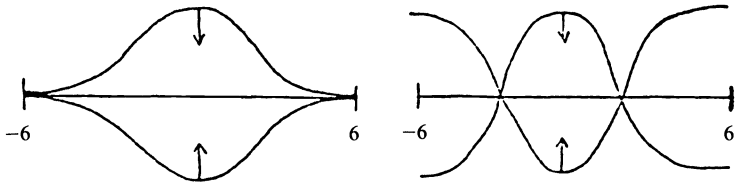


FIG. 5. *Comparison of a nonlinear K_1 breather mode (left) versus a linear K_1 Fourier mode (right).*

Thus, we view this ode system as a model problem and a preview of [5]. Moreover, in [11] we discuss the inclusion of the K_2 Fourier mode into a three-complex mode truncation.

Remark 2. However, this ansatz is capable of modeling some of the apparent features of the perturbed sine-Gordon structure as discussed in § 1. In particular, recall that the chaotic sine-Gordon dynamics (at $\varepsilon\Gamma = .103$, Fig. 3) reflect a competition between a discrete number (here two $K_0 \oplus K_1$) of weakly unstable structures: one with the breather in the center, the other with the breather localized at the ends of the interval. These pde states are, equivalently, related by a half-period translation. Moreover, the pde flow from one state to the other is through the flat K_0 state, which has associated homoclinic orbits in the unperturbed phase space, whereas the unperturbed $K_0 \oplus K_1$ states are neutrally stable in the absence of the perturbation and develop weak instabilities at the onset of chaos.

Now consider the perturbed sine-Gordon solution u^ε, as modeled by this two mode ansatz:

$$u^\varepsilon \sim 2\sqrt{\varepsilon\tilde{\omega}}[(c(T)+b(T)\cos(kX))\, e^{i\omega t}+c.c.]+O(\varepsilon\tilde{\omega}), \tag{2.6}$$

with $c(T)$, $b(T)$ governed by (2.5). These perturbed odes (2.5) admit the symmetry $(c, b)\to(c, -b)$, which is equivalent in u^ε to a translation by $L/2$. Moreover, this symmetry implies that $b = 0$ is an invariant subspace, which for u^ε in (2.6) corresponds to the flat intermediate structure. (The equivalences extend further as discussed in § 3.)

Remark 3. The truncated ansatz (2.6), with c, b governed by (2.5), is robust enough to capture all three spectral configurations of Fig. 3, the $K_0 \oplus K_1$ "gap" state, the $K_0 \oplus K_1$ "cross" state, and the intermediate K_0 state with complex double points and associated homoclinic components, as we will see below. Therefore, this approximation has the potential to flow between gap and cross ($K_0 \oplus K_1$) spectral configurations by passing through the homoclinic K_0 configuration. Recall from Fig. 2(b) and Remark 2 above that this is the spectral flow of the perturbed sine-Gordon equation.

3. Properties of the unperturbed modal equations. In our studies of the perturbed sine-Gordon equation [1], [2], [5], we consistently aim to interpret the perturbed system by projection into the phase space of the integrable sine-Gordon equation. Our understanding of finite-dimensional invariant sets in the exact phase space is the foundation of our studies of the perturbed problem. Consistent with this philosophy, we now describe properties of the unperturbed modal system:

$$\begin{aligned} &-ic_T+(|c|^2+\tfrac{1}{2}|b|^2-1)c+\tfrac{1}{2}(cb^*+c^*b)b=0,\\ &-ib_T+(|c|^2+\tfrac{3}{4}|b|^2-(1+k^2))b+(cb^*+c^*b)c=0. \end{aligned} \tag{3.1}$$

Property 1 (integrable Hamiltonian structure). The two complex- (four real-) dimensional system (3.1) is an integrable Hamiltonian system, with two real independent integrals:

$$\begin{aligned} &I=|c|^2+\tfrac{1}{2}|b|^2,\\ &H=\tfrac{1}{2}|c|^2+|b|^2|c|^2+\tfrac{3}{16}|b|^4-\tfrac{1}{2}(1+k^2)|b|^2-|c|^2+\tfrac{1}{4}(b^2c^{*2}+b^{*2}c^2). \end{aligned} \tag{3.2}$$

The system (3.1) can be placed in complex Hamiltonian form as follows. Let $q_1 = c$, $p_1 = c^*$, $q_2 = b/\sqrt{2}$, $p_2 = b^*/\sqrt{2}$, so that the "energy" H takes the form

$$H(q_1, q_2, p_1, p_2)=\tfrac{1}{2}q_1^2p_1^2+2q_1q_2p_1p_2+\tfrac{3}{4}q_2^2p_2^2-(1+k^2)q_2q_2-q_1p_1+\tfrac{1}{2}(q_2^2p_1^2+q_1^2p_2^2).$$

Then Hamilton's equations with this complex structure are

$$\begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix} = i \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix} \nabla_{\begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix}} H,$$

which are precisely equations (3.1) and their complex conjugates.

Property 2 (symmetries). The integrable odes (2.1) admit the following symmetries:

(3.3a) (i) $(c, b) \to (-c, b)$,

(3.3b) (ii) $(c, b) \to (c, -b)$,

(3.3c) (iii) $(c, b) \to (e^{i\phi} c, e^{i\phi} b)$ for any $\phi \in R$.

The reflection symmetries (i), (ii) yield two invariant planes: $c = 0$ and $b = 0$. The S^1 symmetry (iii) yields a circle of fixed points for each nontrivial fixed point.

Property 3 (sets of fixed points). The system (3.1) has three rings of fixed points:

(3.4a) *Ring* 1. $(c, b) = (e^{i\phi}, 0)$, $\phi \in [0, 2\pi)$ in the $b = 0$ invariant subspace;

(3.4b) Ring 2. $(c, b) = \left(0, e^{i\phi} \sqrt{\frac{4}{3}(1+k^2)}\right)$, $\phi \in [0, 2\pi)$, in the $c = 0$ invariant subspace;

(3.4c) *Ring* 3. $(c, b) = \left(e^{i\phi} \sqrt{\frac{1+2k^2}{5}}, 2e^{i\phi} \sqrt{\frac{2-k^2}{15}}\right)$, $\phi \in [0, 2\pi)$.

Remark. For larger lengths $L > 12$, another fixed-point ring exists [11].

The quadrature solution of these integrable equations is most easily affected by the polar coordinate form of (3.1).

Property 4 (polar form of the unperturbed odes). Let $c = C e^{i\gamma}$, $b = B e^{i\beta}$, $\theta = 2(\gamma - \beta)$; then (3.1) becomes

$$\begin{aligned} &C_T + \tfrac{1}{2} C B^2 \sin\theta = 0, \qquad B_T - C^2 B \sin\theta = 0, \\ &\theta_T + (I + 2k^2) - 3C^2 + 2(I - 2C^2)\cos\theta = 0. \end{aligned} \tag{3.5}$$

For completeness, we also list

$$\begin{aligned} &\gamma_T + C^2 - 1 + \tfrac{1}{2} B^2 (2 + \cos\theta) = 0, \\ &\beta_T + \tfrac{3}{4} B^2 + C^2 (2 + \cos\theta) - (1 + k^2) = 0. \end{aligned}$$

By use of the integrals I and H, it is now easy from (3.5) to effect a complete reduction to quadrature solutions of each choice $I = I_0$, $H = H_0$. These general formulas are not the focus of this paper but will be presented elsewhere [4]. Some special cases will be relevant here (see Property 6 below).

Property 5 (stability type for each ring of fixed points). The S^1 symmetry, together with the fact that the amplitudes are constant when $B = 0$ (Ring 1), or $C = 0$ (Ring 2), or $\theta = 0$ (Ring 3), produce a double linearized eigenvalue of zero for each ring. The linearized stability of these fixed points is therefore straightforward to compute; we find:

Ring 1. $(c, b) = (e^{i\phi}, 0)$ has a double zero eigenvalue with associated eigenvectors in the $b = 0$ subspace, and a one-dimensional stable and unstable eigenspace, with $O(1)$ stable and unstable eigenvalues, $\pm k\sqrt{2 - k^2} \approx 1$.

Ring 2. $(c, b) = (0, e^{i\phi}\sqrt{\frac{4}{3}(1+k^2)})$ has a double zero eigenvalue with corresponding eigenvectors in the $c = 0$ subspace, and two purely imaginary, complex conjugate eigenvalues, $\pm i\sqrt{\frac{1}{3}(4k^4-1)}$. These fixed points are purely center-like.

Ring 3.

$$(c, b) = \left(e^{i\phi}\sqrt{\frac{1+2k^2}{5}}, 2\, e^{i\phi}\sqrt{\frac{2-k^2}{15}} \right)$$

has a double zero eigenvalue with corresponding eigenvectors in the $\theta = 0$ subspace, and two purely imaginary eigenvalues $\pm i\sqrt{7}|c||b|$. These fixed points are centers.

Property 6 (homoclinic orbits associated to fixed-point Ring 1). The unstable fixed points on Ring 1, with $c = e^{i\phi}$, $b = 0$, lie on the energy surface $H = -\frac{1}{2}$, $I = 1$. These are the asymptotic states associated to heteroclinic orbits on this energy surface, which correspond to the one-dimensional stable and unstable manifolds of each fixed point on Ring 1.

Using the integrals I and H from (3.2), we find a convenient integral is

$$h = H - (\tfrac{1}{2}I^2 - I),$$

which can be manipulated to find

$$\cos\theta = \frac{2h + B^2(\frac{3}{8}B^2 + k^2 - I)}{B^2(I - \frac{1}{2}B^2)}.$$

Using this formula, the polar equations (3.5), and the values $H = -\frac{1}{2}$, $I = 1$, $h = 0$ appropriate to $(c = e^{i\phi}, b = 0)$, we determine an effective oscillator equation for $z = B^2$:

$$\frac{1}{2}z_T^2 + \left[-\frac{z^2}{32}(z - 8k^2)(7z - 8(2-k^2)) \right] = h = 0.$$

The familiar potential energy diagram below (Fig. 6), with energy level $h = 0$, exhibits the infinite-period behavior of $z = B^2 : 0 \nearrow \sqrt{\frac{8}{7}(2-k^2)} \searrow 0$, where $\nearrow$, $\searrow$ denote monotonically increasing, decreasing behavior, respectively.

The remaining formulas for C, γ, β are similarly derived [4]. There are also additional orbits homoclinic to the closed curves nested around Ring 1 in the $b = 0$ invariant subspace (see Property 8).

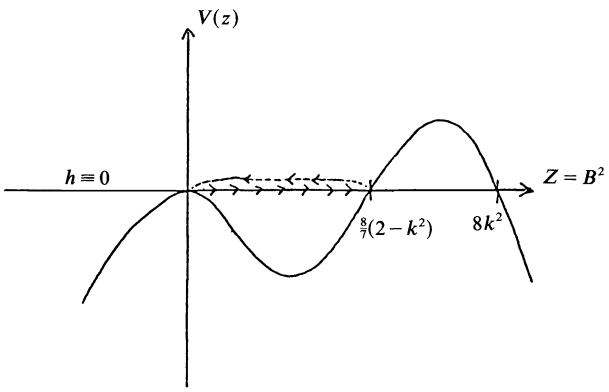


FIG. 6. *Potential energy diagram for* $z = B^2 = |b|^2$.

Property 7 (connection between the ode fixed points and sine-Gordon solutions). In the asymptotic representation (2.6) of sine-Gordon solutions, the above fixed points of the unperturbed odes reflect the following solutions and their stability properties.

Ring 1. $u \sim 2\sqrt{\varepsilon\tilde{\omega}}[c\, e^{i\omega t} + c^*\, e^{-i\omega t}]$, which corresponds to the K_0 flat pendulum solution, frequency locked to the driver frequency ω. The $O(1)$ instability of these fixed points in the unperturbed system reflects the $O(1)$ instability of the exact K_0 sine-Gordon solution (recall Fig. 3(a) and the surrounding discussion). Moreover, the orbits homoclinic to Ring 1 reflect the sine-Gordon solutions which are homoclinic to the pendulum solution with frequency $\omega = .87$.

Ring 2. $u \sim \sqrt{\varepsilon}[b \cos kX\, e^{i\omega t} + c.c.]$ corresponds to the pure K_1 mode, with a zero-mean (K_0) component, and with frequency of the driver. These solutions exist for sine-Gordon, but are not observed in the perturbed dynamics. This is presumably explained by the larger amplitude of this Ring 2, $|b| \approx \sqrt{8/3}$, relative to the other Rings 1, 3. These corresponding solutions in the perturbed pde would then be expected to show up by varying initial conditions with a significantly larger energy, or by driving the system harder (see § 4).

Ring 3. $u \sim \sqrt{\varepsilon}[(c + b \cos kX)\, e^{i\omega t} + c.c.]$ corresponds to the $K_0 \oplus K_1$ sine-Gordon solution, consisting of the K_1 breather plus nonzero mean, locked at the frequency ω. These states are observed in the perturbed dynamics. Moreover, the unperturbed pde stability type (neutrally stable) agrees with that of the unperturbed odes.

In summary, the fixed-point Rings 1 and 3 in the unperturbed odes reflect remarkably well the unperturbed K_0 and $K_0 \oplus K_1$ sine-Gordon solutions, and moreover maintain a parallel linearized instability and homoclinic orbit structure of the K_0 state, as well as the neutral stability of the $K_0 \oplus K_1$ states.

Below (§ 4) we discuss how the perturbation selects individual points from these rings.

Property 8 (simple periodic orbits nested around the fixed-point rings). The symmetries of the unperturbed odes, Property 2, lead to a nesting of closed curves in the subspaces containing Rings 1, 2, and 3. For example, in the invariant subspace $b = 0$, we find closed curves $|C| = \text{constant} = C_0$, which yields the one-parameter family of periodic solutions:

$$c = C_0\, e^{i(1-C_0^2)T} \quad C_0 = \text{constant}, \qquad b = 0.$$

As $C_0 \to 1$, these curves approach Ring 1 while the frequency goes to zero.

Similarly, in the $C = 0$ invariant subspace there is a one-parameter family of periodic solutions surrounding Ring 2:

$$c = 0, \qquad b = B_0 \exp\left(i(1 + k^2 - \tfrac{3}{4}B_0^2)T\right).$$

In the $\theta = 0$ subspace, which contains Ring 3, we find another one-parameter family of closed curves corresponding to periodic solutions:

$$c = \sqrt{\tfrac{3}{8}B_0^2 + \tfrac{1}{2}k^2}\, \exp\left(i(1 - \tfrac{1}{2}k^2 - \tfrac{15}{8}B_0^2)T\right),$$
$$b = B_0 \exp\left(i(1 - \tfrac{1}{2}k^2 - \tfrac{15}{8}B_0^2)T\right).$$

Note that these nested closed curves around Rings 1 and 3 are connected at the periodic solution $b = 0$, $c = (k/\sqrt{2}) \exp\left(i(1 - \tfrac{1}{2}k^2)T\right)$. The Floquet stability analysis of these one-parameter families of periodic solutions yields coupled Mathieu equations, to be discussed elsewhere.

4. Bifurcations of the perturbed modal equations. We now discuss how the above properties of the integrable odes (3.1) reveal themselves in the bifurcation structure and dynamics of equations (2.5). Recall that we fix $\tilde{\alpha} = .155$, and consider the bifurcations of (2.5) as the constant driver $\tilde{\Gamma}$ is varied.

The following bifurcation curves (Fig. 7) were originally generated for us by Jolly and Kevrekides using the code AUTO, and since then verified by us, while the linearized eigenvalues and associated eigenvectors were independently computed along the curves by Hyman using the CLAMS package.

We now discuss these bifurcation curves and the associated dynamics.

Property 1 (existence of fixed points as a function of $\tilde{\Gamma}$). Branch *OABFG* is a pure K_0 branch, consisting of steady states with $b = 0$.

Branch *BCD* is a double $K_0 \oplus K_1$ branch, consisting of fixed points (c, b) and $(c, -b)$, with $b \neq 0$, $c \neq 0$. (Recall from Remark 2 in § 2 that the perturbed system (2.5) retains the reflection symmetry $(c, b) \to (c, -b)$, so that all fixed points with $b \neq 0$ come in pairs with equal l_2 norm.) The bifurcation point *B* corresponds both to the change of stability of the K_0 branch of fixed points from one to two unstable dimensions and to the beginning of the $K_0 \oplus K_1$ branch of fixed points.

Some features of this bifurcation diagram are found either analytically or by simple perturbation theory arguments, as we now sketch.

Property 2 (explicit parameterization of the entire K_0 branch of fixed points). With $b = 0$, the fixed points of (2.5) satisfy, with $c = c_1 + ic_2$, $\tilde{\alpha} \approx .155$,

$$\begin{aligned} (c_1^2 + c_2^2 - 1)c_1 + \tilde{\alpha}c_2 &= 0, \\ (c_1^2 + c_2^2 - 1)c_2 - \tilde{\alpha}c_1 &= \tilde{\Gamma}. \end{aligned} \tag{4.1}$$

If we fix $n = \sqrt{c_1^2 + c_2^2} = l_2$ norm of $(c, 0)$, the equations (4.1) represent two orthogonal

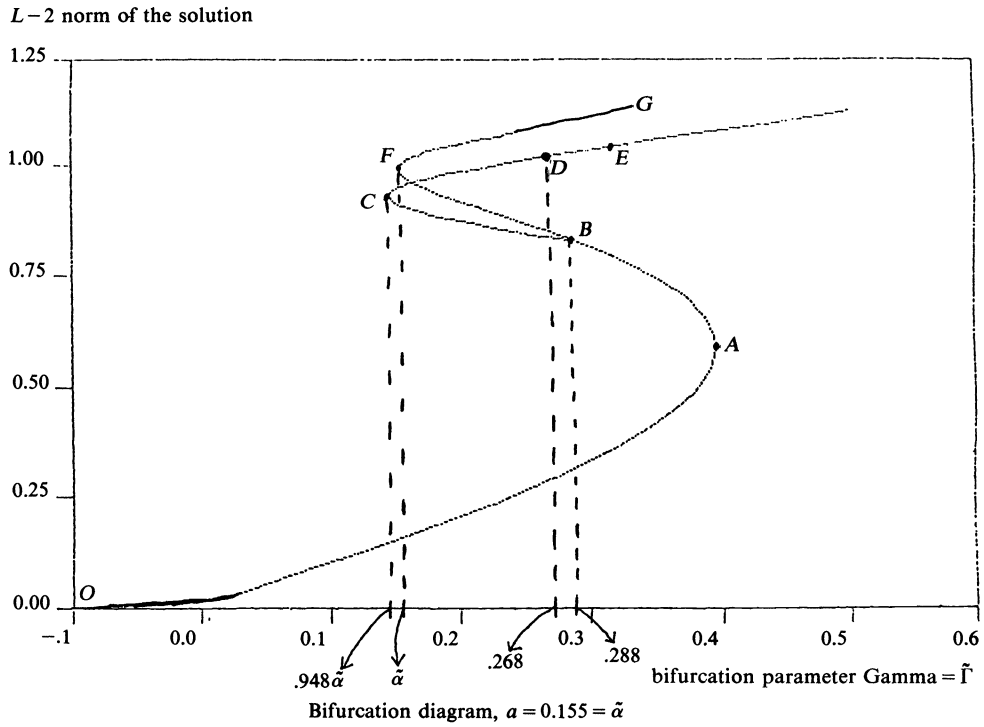


FIG. 7. *Bifurcation diagram of* (2.5) *as* $\tilde{\Gamma}$ *is varied.*

lines in the (c_1^2, c_2^2) plane. We then pick $\tilde{\Gamma}$ such that the two lines intersect on the circle of radius n. This algorithm yields the inverse of the K_0 curve of Fig. 7, which is a nice function:

$$\tilde{\Gamma}(n) = n\sqrt{\tilde{\alpha}^2 + (1-n^2)^2}. \tag{4.2}$$

In this way, we analytically generate the K_0 curve $OABFG$, and calculate the turning points A and F by setting $\tilde{\Gamma}'(n) = 0$, which verifies the numerically generated curve.

Property 3 (perturbation calculations to describe the turning points in the bifurcation branches). By a classical perturbative phase-locking condition, we reproduce the qualitative multibranch structure of Fig. 7, identify these fixed-point branches as phase-locked fixed points from the unperturbed fixed-point Rings 1, 3, and quantitatively capture the turning points or "bends" in the bifurcation curves.

First we compute how the unperturbed integral I varies in the presence of the perturbed dynamics (2.5) (recall that $\tilde{\alpha}$ and $\tilde{\Gamma}$ are our rescaled small parameters):

$$\begin{aligned} \frac{dI}{dT} &= -2\tilde{\alpha}\left[|c|^2 + \frac{1}{2}|b|^2\right] - 2\tilde{\Gamma}\,\mathrm{Re}\,(c) \\ &= -2\tilde{\alpha}I - 2\tilde{\Gamma}\,\mathrm{Re}\,(c). \end{aligned} \tag{4.3}$$

Next we seek fixed points which are perturbations of Rings 1, 2, 3 in (3.4a)-(3.4c) evaluate (4.3) on this ansatz, and demand that dI/dT vanishes to $O(\tilde{\alpha}, \tilde{\Gamma})$—which selects the phase(s) of Rings 1, 2, 3 that "lock(s)" to the perturbation. This procedure yields the following nonresonance or phase-locking conditions:

Ring 1. With $c = e^{i\phi} + \varepsilon\Delta c$, $b = \varepsilon\Delta b$, $0 < \varepsilon \ll 1$, phase-locking criterion: $\tilde{\alpha} + \tilde{\Gamma}\cos\phi = 0$.

Ring 2. With $c = \varepsilon\Delta c$, $b = e^{i\phi}\sqrt{\frac{4}{3}}(1+k^2) + \varepsilon\Delta b$, phase-locking criterion: $\tilde{\alpha}I_0 + \varepsilon\tilde{\Gamma}\,\mathrm{RE}\,(\Delta c) = 0$, where RE ( ) ≡ real part ( ).

Ring 3. With

$$c = e^{i\phi}\sqrt{\frac{1+2k^2}{5}} + \varepsilon\Delta c, \qquad b = 2\,e^{i\phi}\sqrt{\frac{2-k^2}{15}} + \varepsilon\Delta b,$$

phase-locking criterion:

$$\tilde{\alpha}\left(\frac{7+4k^2}{15}\right) + \tilde{\Gamma}^2\cos\phi\sqrt{\frac{1+2k^2}{5}} = 0.$$

This perturbation analysis yields the following conclusions (recall $\tilde{\alpha} = .155$):

Ring 1. $\cos\phi = -\tilde{\alpha}/\tilde{\Gamma}$, so this ring does not phase lock until $\tilde{\Gamma} \geqq \tilde{\alpha}$ (which precisely yields the turning point F), and for $\tilde{\Gamma} > \tilde{\alpha}$ exactly two phases are selected, corresponding to the two branches FG and FB. (The stability of these and other branches is discussed in the next property.)

Ring 2. If $\tilde{\alpha}$, $\tilde{\gamma} = O(\varepsilon)$, there are no solutions of the nonresonance condition ($\tilde{\alpha}I_0 \neq 0$). However, if $\tilde{\alpha} = O(\varepsilon^2)$, $\tilde{\Gamma} = O(\varepsilon)$, then we find a balance in this equation. This occurs when $\tilde{\Gamma} = O(\sqrt{\tilde{\alpha}})$, which is outside the range of our diagram and so will not be of interest here. (Referring back to § 3, Property 7, Ring 2, we now find these zero-mean solutions do not resonate with the perturbation until the system is driven harder.)

Ring 3. The phase-locking condition yields

$$\cos\phi = -\frac{\tilde{\alpha}}{\tilde{\Gamma}}\frac{7+4k^2}{15}\cdot\sqrt{\frac{5}{1+2k^2}},$$

which correctly predicts the turning point C and the two emanating branches.

Property 4 (stability of the bifurcation branches). In the phase-locked branches of Fig. 7, the upper branches *FG* and *CD* locally inherit the stability type of the unperturbed Rings 1 and 3, respectively, while the lower branches *FB* and *CB* locally pick up an additional weak unstable eigenvalue due to the perturbation. (These facts are easily deduced perturbatively.)

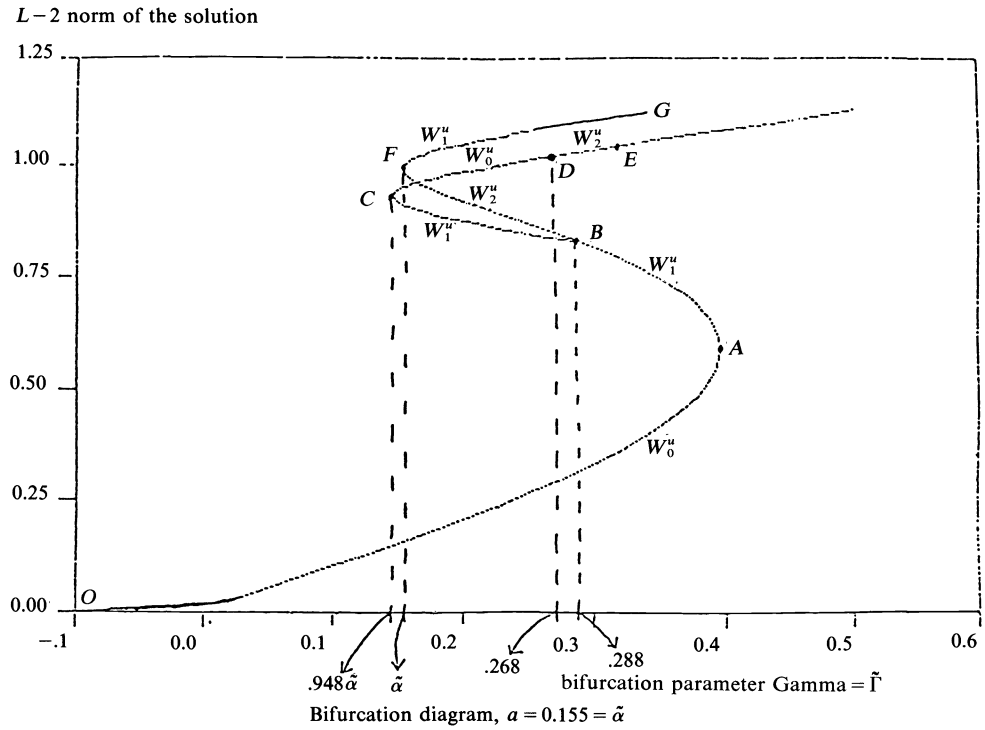
The stability type of all branches in Fig. 7 is numerically computed, with the following results. Let W^s_k, W^u_k denote a k-dimensional stable or unstable manifold, respectively. Then the following diagram indicates the stability type of each branch: (recall that in dimension four, $W^s_k(\bar{X}_0)$, $W^u_l(\bar{X}_0)$ for hyperbolic fixed points satisfy $k+l=4$, so it suffices to list $W^u_k(\bar{X}_0)$):

Highlights. (1) The K_0 branch *FG* is the phase-locked continuation branch of Ring 1, which maintains the one-dimensional unstable manifold character of the homoclinic orbits to Ring 1 in the unperturbed problem. The *FG* branch, therefore, is the perturbed ode signature of the homoclinic pde structures described earlier in § 1.

(2) The double $K_0 \oplus K_1$ branch *CD* is the stable phase-locked continuation branch of Ring 3, which corresponds in the perturbed pde to the phase-locked, stable, breather plus nonzero mean $(K_0 \oplus K_1)$ solutions.

(3) The point *D* on the $K_0 \oplus K_1$ branch corresponds to the subcritical Hopf bifurcation. At this value of $\tilde{\Gamma} \approx .268$, the previously stable $K_0 \oplus K_1$ breather plus mean solutions become two-dimensionally, weakly unstable due to the perturbation. As we discuss below, just after this Hopf bifurcation the perturbed system goes chaotic.

Since this Hopf bifurcation is subcritical, the associated periodic orbits of this bifurcation phenomenon are unstable, and these are not observed in our numerical simulations. This fact is quite consistent with the pde bifurcation structure (Fig. 1 and remarks just below it), where in this parameter regime we did not see quasi periodicity



Bifurcation diagram, $a = 0.155 = \tilde{\alpha}$

FIG. 8. *Bifurcation diagram with stability type of each branch of fixed points.*

prior to chaos. This model suggests that when the second frequency is generated by Hopf bifurcation in the pde, it is unstable in this specific parameter regime.

Property 5 (global connections between fixed points: before and after the Hopf bifurcation). We have numerically determined the global connections of the unstable manifolds of fixed points at given stress parameter values $\tilde{\Gamma}$. These fixed-point connections were numerically computed in two ways, first with the package UMFUT of Doedel. We then independently verified the connections using the ode package LSODE. The method we use is to initialize the ode system with the coordinates of the unstable fixed points plus a small $(0(10^{-2}-10^{-3}))$ increment in the direction of the unstable eigenvector(s). The orbit then converges to the indicated fixed point. The only numerically sensitive connections are saddle-saddle connections, for which we impose tighter error bounds on the ode code and more careful resolution of the unstable manifold direction(s) to select initial conditions for the connecting orbit. The saddle-saddle connections are confirmed by restriction to the invariant subspace $b=0$, where the connections become saddle-stable node, which are numerically stable.

The interesting and relevant connections for this discussion are those for $\tilde{\Gamma}$ preceding and following the Hopf bifurcation at $\tilde{\Gamma}\approx .268$. We indicate these schematically in Fig. 9.

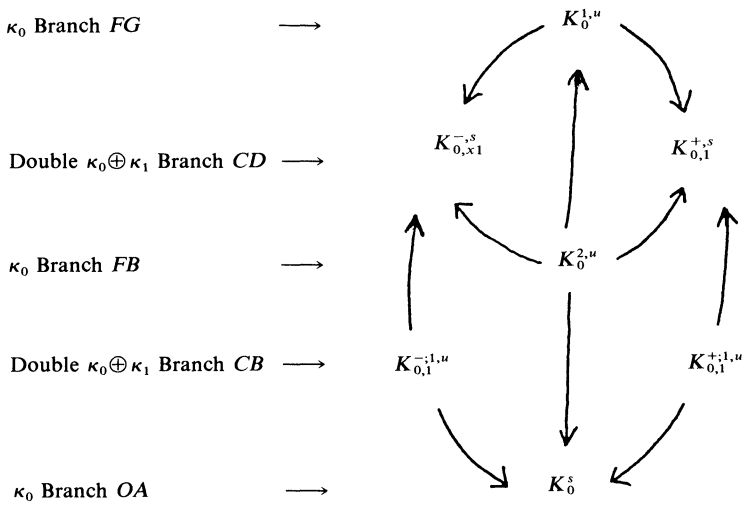


FIG. 9. *Global fixed-point connections* $.155<\tilde{\Gamma}<.268$.

These connections are quite expected from the pde. For example, the large amplitude unstable mean $(K_0^{1,u})$ is unstable to the nearest energy stable state, which is either the stable breather plus mean $(K_{0,1}^{+,s})$ or its equal energy translate $(K_{0,1}^{-,s})$. Thus, the unstable manifold of $K_0^{1,u}$ lands in one direction at $K_{0,1}^{+,s}$, in the other at $K_{0,1}^{-,s}$. Another example is the unstable breather plus mean $K_{0,1}^{+;1,u}$. Along perturbations which decrease energy, the state is unstable to the stable flat configuration (K_0^s) since there is not enough energy to sustain the spatial structure. If energy is increased, however, the state is unstable to the higher energy breather plus mean $K_{0,1}^{+,s}$, which is stable.

After the Hopf bifurcation $\tilde{\Gamma}\gtrsim .268$, all that changes in Fig. 9 is that the stable $K_{0,1}^{\pm;s}$ fixed points become two-dimensionally, weakly unstable,

$$K_{0,1}^{\pm;s} \xrightarrow[\text{Hopf}]{\text{after}} K_{0,1}^{\pm;2,u}.$$

We observe numerically, however, that the unstable trajectories out of $K_0^{1,u}$, $K_0^{2,u}$, and $K_0^{\pm;1,u}$ (which previously, Fig. 9, converges to $K_{0,1}^{\pm,s}$) now flow very near to $K_{0,1}^{+;2,u}$ or $K_{0,1}^{-;2,u}$, dominated by the $O(1)$ attracting directions (stable dimension two), but slowly build up the weak instabilities, leave this neighborhood, then quickly land back in a neighborhood of either $K_{0,1}^{+;2,u}$ or $K_{0,1}^{-;2,u}$. The dynamics of these odes depicts classic intermittent chaos: the "laminar" behavior is characterized by settling into a neighborhood of $K_{0,1}^{+;2,u}$ or $K_{0,1}^{-;2,u}$, and the intermittent chaotic bursts are associated with the flight out of the neighborhood of one and subsequent landing back into either of them.

In Fig. 10 we exhibit the time evolution for a selected orbit of the perturbed odes (2.5) at $\tilde{\Gamma} = .275$, which corresponds exactly to the $\varepsilon\Gamma = .103$ pde numerical experiments shown in § 1, Figs. 2 and 3. The initial condition of Fig. 10 is taken along the unstable manifold nearby $K_{0,1}^{+;2,u}$, which is directly into the conjectured strange attractor. For consistency, we also continue this numerical run to $t = 10{,}000$ (corresponding to $t \sim 70{,}000$ in the pde time units), and the run remained chaotic. The leading Lyapunov number was computed to be $2^{.17}$.

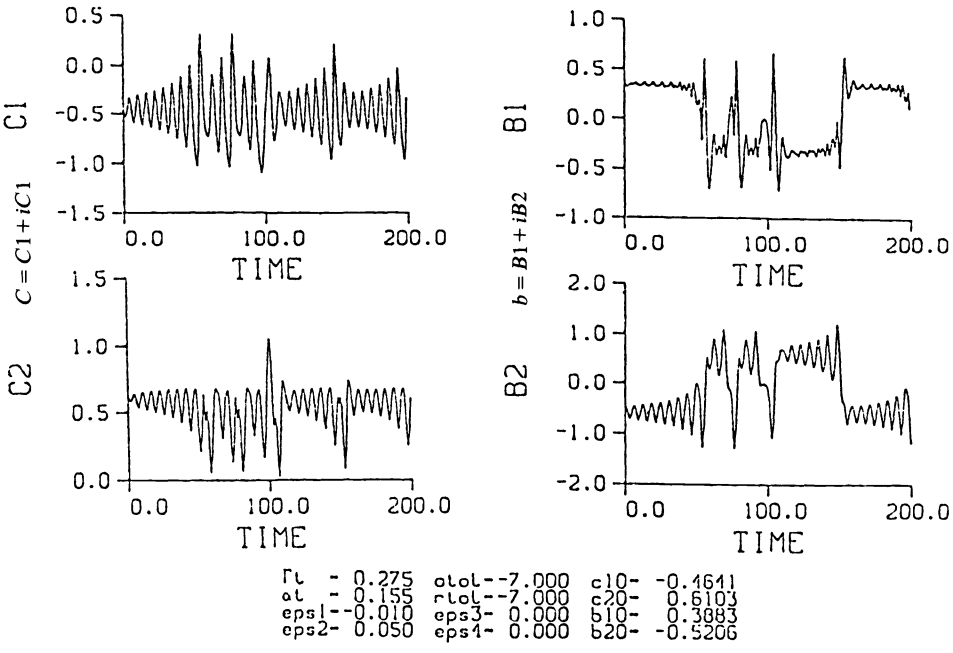
Motivated by our pde study in § 1 and the identification of homoclinic K_0 structures in the chaotic dynamics, we now seek the analogue of homoclinic crossings in these model odes. We have so far identified a parallel structure between this model problem and the pde: the upper $K_0^{1,u}$ fixed point corresponds to the K_0, order 1 unstable flat state, associated to the unperturbed homoclinic components; the $K_{0,1}^{\pm;2,u}$ fixed points correspond to breather plus nonzero mean states, phase shifted by a half period, which are neutrally stable in the unperturbed equation, but which develop weak instabilities due to the perturbation when $\tilde{\Gamma} \gtrsim .268$.

The perturbed pde as evidenced in Figs. 2(a) and 2(b), exhibits intermittent chaos characterized by a passage out of a "laminar" $K_0 \oplus K_1$ state, through the homoclinic K_0 state, and then back into another weakly unstable $K_0 \oplus K_1$ state. In this perturbed ode, this behavior corresponds to the schematic loop of Fig. 11.

In summary, our numerical studies of the ode and pde clearly suggest that the onset of observable chaos may be described by a jumping process between two weakly unstable coherent states. (These two states are related by a discrete symmetry, $(c, b) \to (c, -b)$ in the ode, and in the pde by a half-period translation from a state localized in the center or the ends of the spatial interval.) In the *unperturbed ode and pde systems*, homoclinic orbits have been identified which are homoclinic in the pde to x-independent, order 1 unstable periodic solutions while in the ode these orbits are homoclinic to the ring of fixed points, $|c| = 1$, $b = 0$. In each system, these orbits are homoclinic to degenerate solutions that intermediate the unperturbed spatially localized solutions. *Moreover*, in both the perturbed ode and pde systems, we have numerically correlated the jumping process with unperturbed homoclinic crossings.

It is therefore clear to us that a Melnikov-type calculation is appropriate, centered on the unperturbed homoclinic orbits. The goal of this analysis is to establish our conjecture for the observable chaotic dynamics: the existence of horseshoes in the perturbed dynamics, which rigorously identifies the jumping process in the ode and pde as topologically conjugate to a Bernoulli shift on two symbols. (The two symbols represent the states localized in the center and wings of the interval.)

A precise dynamical systems mechanism for the observable chaos has been formulated in collaboration with Wiggins and Kovacic. The ode scenario is based on existence of a four-dimensional Silnikov-like structure (Guckenheimer and Holmes [13, § 6.5], and Wiggins [14, § 3.2]); the rigorous proof is in progress by Wiggins and Kovacic and will appear in the thesis of Kovacic. The extension of this rigorous analysis to the pde is in progress [15].



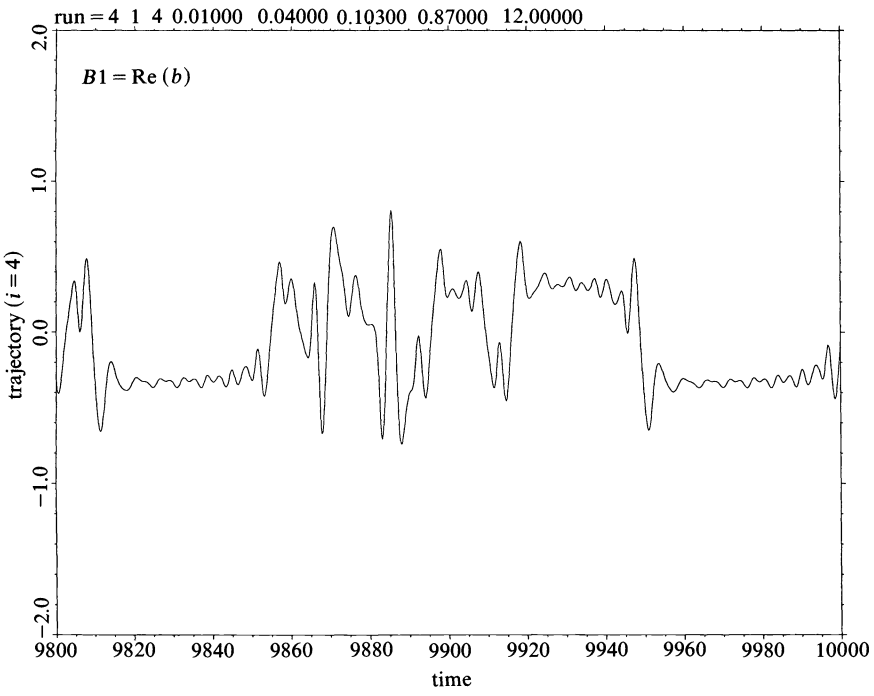


FIG. 10. *Time evolution of the odes* (2.5) *in the chaotic regime at* $\tilde{\Gamma} = .275$, *corresponding to the pde experiments of Figs.* 2 *and* 3 *at* $\varepsilon\Gamma = .103$. *The initial condition for this time series is a small increment* $(0(10^{-2}))$ *from* $K_{0,1}^{+;2,u}$ *in the unstable eigendirections, chosen so that the flow immediately finds the surmised strange attractor. We also include one graph from this time series near* 10,000 *time units, comparable to the pde time series in Fig.* 2.

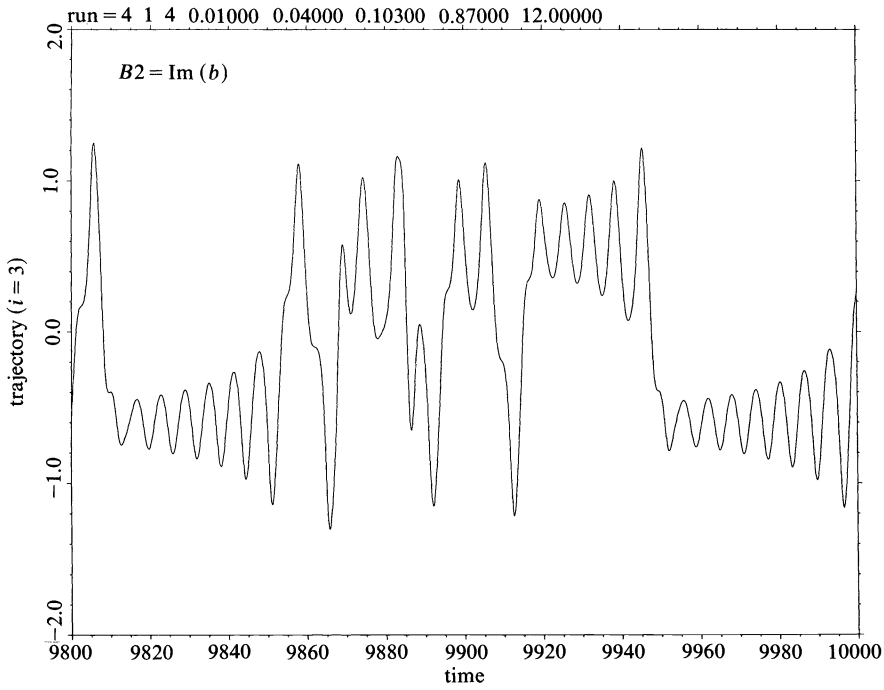


FIG. 10—continued

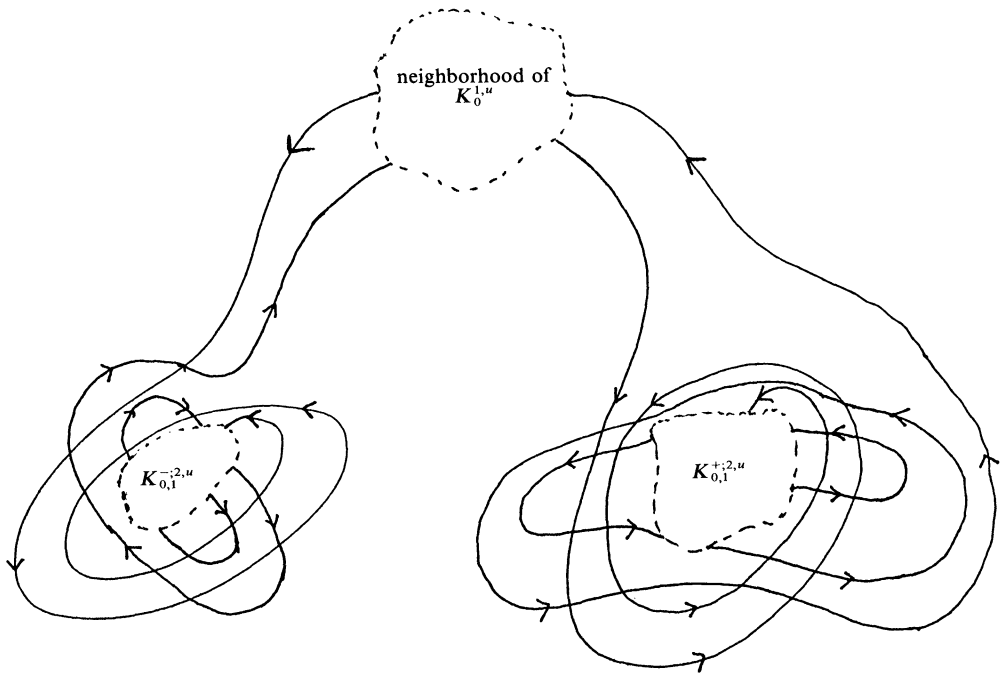


FIG. 11. *Schematic loop of behavior on the chaotic attractor.*

This surmised behavior creates a loop as sketched in Fig. 11. We test this conjecture numerically in the following way.

Starting at $K_{0,1}^{+;2,u}$ or $K_{0,1}^{-;2,u}$, we use the computed unstable eigenvectors to locally span $W_2^u(K_{0,1}^{\pm;2,u})$. We then numerically shoot from these fixed points $K_{0,1}^{\pm;2,u}$ along the unstable manifold, integrate the perturbed flow, and numerically monitor the distance to all fixed points at this value of $\tilde{\Gamma}$. These distance functions are defined as follows relative to the labeling in Fig. 12: DIS j = distance of the computed orbit to fixed point j, etc. These distance functions DIS 1–DIS 7 are provided below (Fig. 13) for one of the representative unstable eigendirections out of $K_{0,1}^{+;2,u}$, which coincides with the time evolution provided in Fig. 10.

Conclusion. The numerical evidence verifies the jumping process between neighborhoods of $K_{0,1}^{+;2,u}$ and $K_{0,1}^{-;2,u}$, with intermediate passages nearby $K_0^{1,u}$ during the jumps. Note that the DIS 2 and DIS 3 functions oscillate near zero as the orbit settles into a neighborhood of $K_{0,1}^{+;2,u}$ or $K_{0,1}^{-;2,u}$, respectively.

$K_0^{1,u} \equiv$ fixed point 1

$K_{0,1}^{+;2,u} \equiv$ fixed point 2 $\qquad$ $K_{0,1}^{-;2,u} \equiv$ fixed point 3

$K_0^{2,u} \equiv$ fixed point 4

$K_{0,1}^{+;1,u} \equiv$ fixed point 5 $\qquad$ $K_{0,1}^{-;1,u} \equiv$ fixed point 6

$K_0^s \equiv$ fixed point 7

FIG. 12. *For* $\tilde{\Gamma} = .275$, *the seven fixed points are assigned a numerical label:* 1 *is assigned to* $K_0^{1,u}$, 2 *is assigned to* $K_{0,1}^{+;2,u}$, *etc.*

Moreover, in the bursts out of these "laminar" states, the orbit gets relatively close to $1 = K_0^{1,u}$ (occasionally very close) as indicated in the graph of DIS 1, whereas the orbit is always $O(1)$ distance from fixed points 4, 5, 6, and 7. Thus, as in the pde simulations of § 1, the perturbed ode apparently goes chaotic coincidentally with random passages through or near the homoclinic structures of the unperturbed problem.

One final measurement of this thesis is the ode analogue of our sine-Gordon spectral projection of the perturbed pde (Fig. 2(b)). The homoclinic unperturbed fixed-point Ring 1 has the integral dependence $H = \frac{1}{2}I^2 - I$, so that $h = H - (\frac{1}{2}I^2 - I) = 0$ on the homoclinic orbit (recall Property 6 of § 3). We now seek to measure the projection of the perturbed flow, relative to this unperturbed homoclinic configuration, by checking for zero crossings of h. The graph of h is provided along with the distance functions in Fig. 13.

5. Correlations between the infinite-dimensional and reduced systems. So far, we have measured homoclinic crossings in two distinct ways: *in the perturbed* pde by graphing the exact sine-Gordon spectrum of u^ε at each timestep, and *in the* ode by graphing $h = H - (\frac{1}{2}I^2 - I)$ and checking for zero crossings. As a final test of this

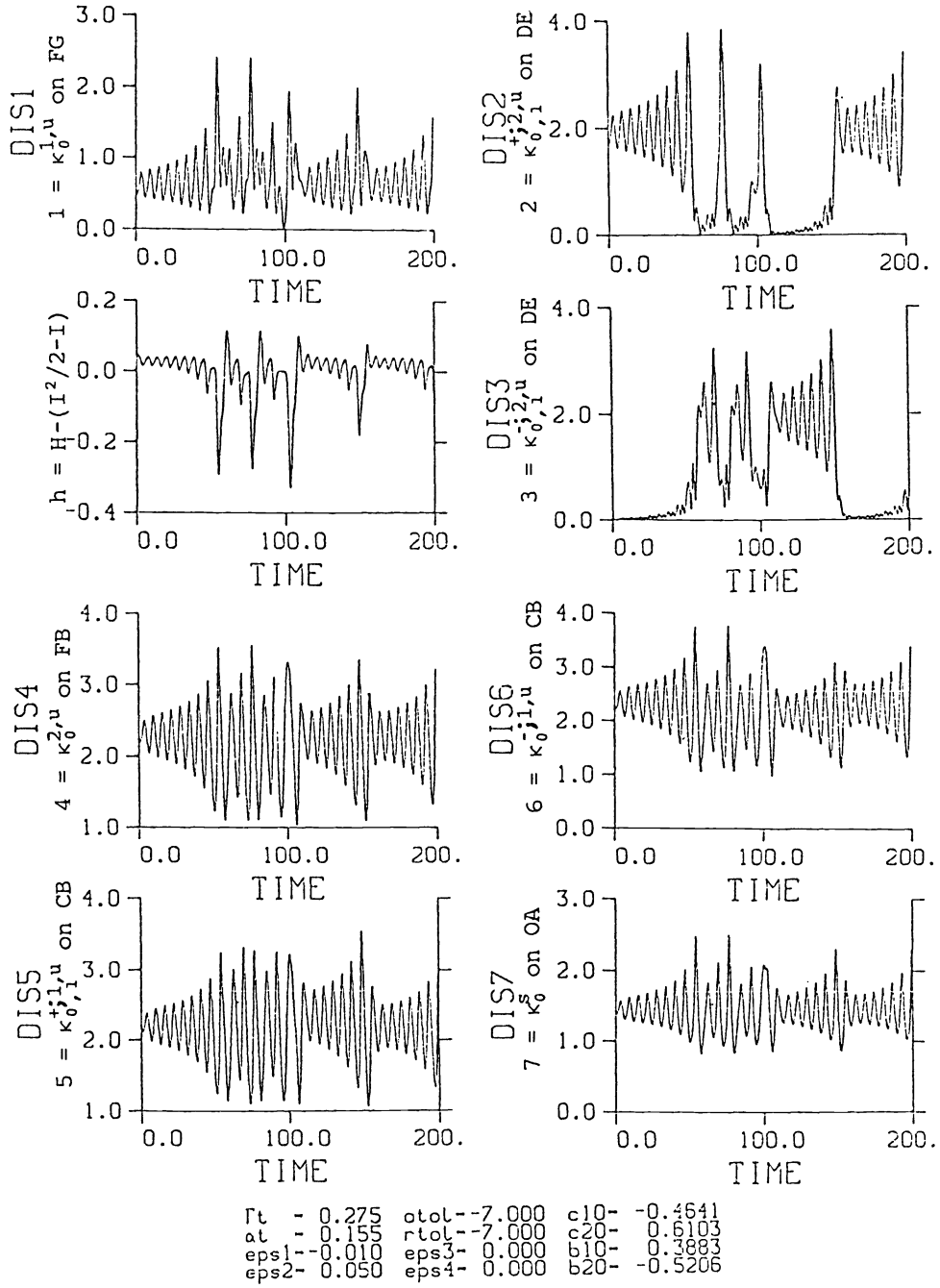


FIG. 13. *With the labeling of fixed points* 1-7 *as in Fig.* 12 (*fixed point* $1 = K_0^{1,u}$, *etc.*), *and with the orbit* $(c(t_n), b(t_n))$ *from Fig.* 10 *on the chaotic attractor, distance functions,* DIS j, $j = 1, \cdots, 7$, *are computed which measure the orbit point's distance to fixed point* j *at each timestep* t_n. *Importantly, note the changes in vertical scales.*

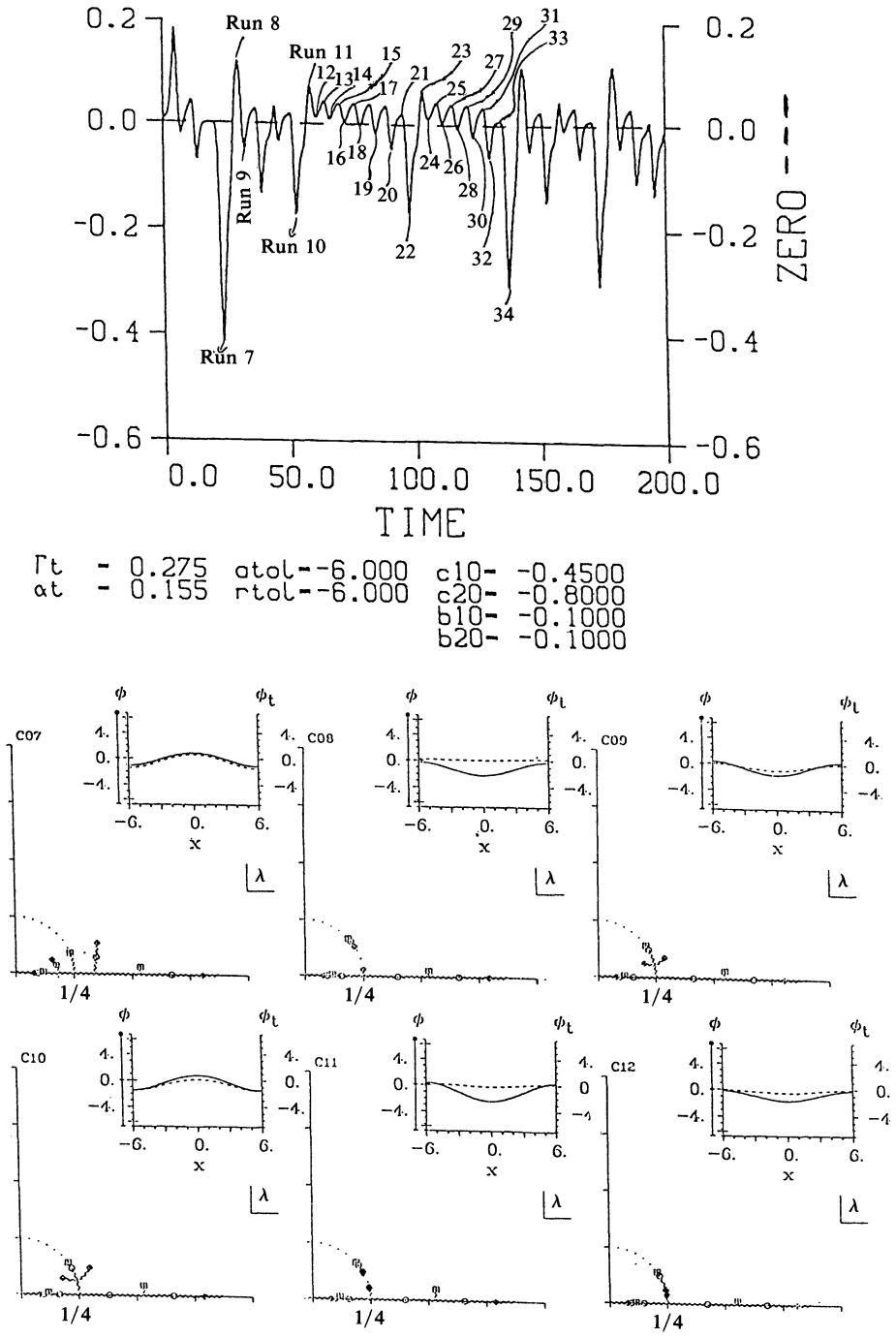


FIG. 14. *The ode and pde homoclinic diagnostics are computed together from an orbit* $(c(t_n), b(t_n))$ *on the chaotic attractor. The top graph is the ode diagnostic,* $h = H - (\frac{1}{2}I^2 - I)$, *which we check for zero crossings, with several discrete times labeled. From the values* $(c(t_n), b(t_n))$ *at these discrete times, the approximate perturbed sine-Gordon solution,* $u^\varepsilon(x, t_n)$, *is computed from formula* (2.1a). *Then the corresponding sine-Gordon spectral measurement of this approximate* u^ε *is computed. We then seek the correlation between zero crossings of the ode diagnostic h and passage through the homoclinic spectral configuration of the pde.*

homoclinic phenomenon, we combine the two measurements. We take $c(T_n)$, $b(T_n)$ during the flow that generates h, reconstruct the perturbed sine-Gordon solution u^ε by the approximation, (2.1a), and then compute the sine-Gordon spectral measurement of u^ε. When h goes through a zero crossing, does the perturbed sine-Gordon field u^ε pass through a homoclinic spectral configuration? The results appear in Fig. 14.

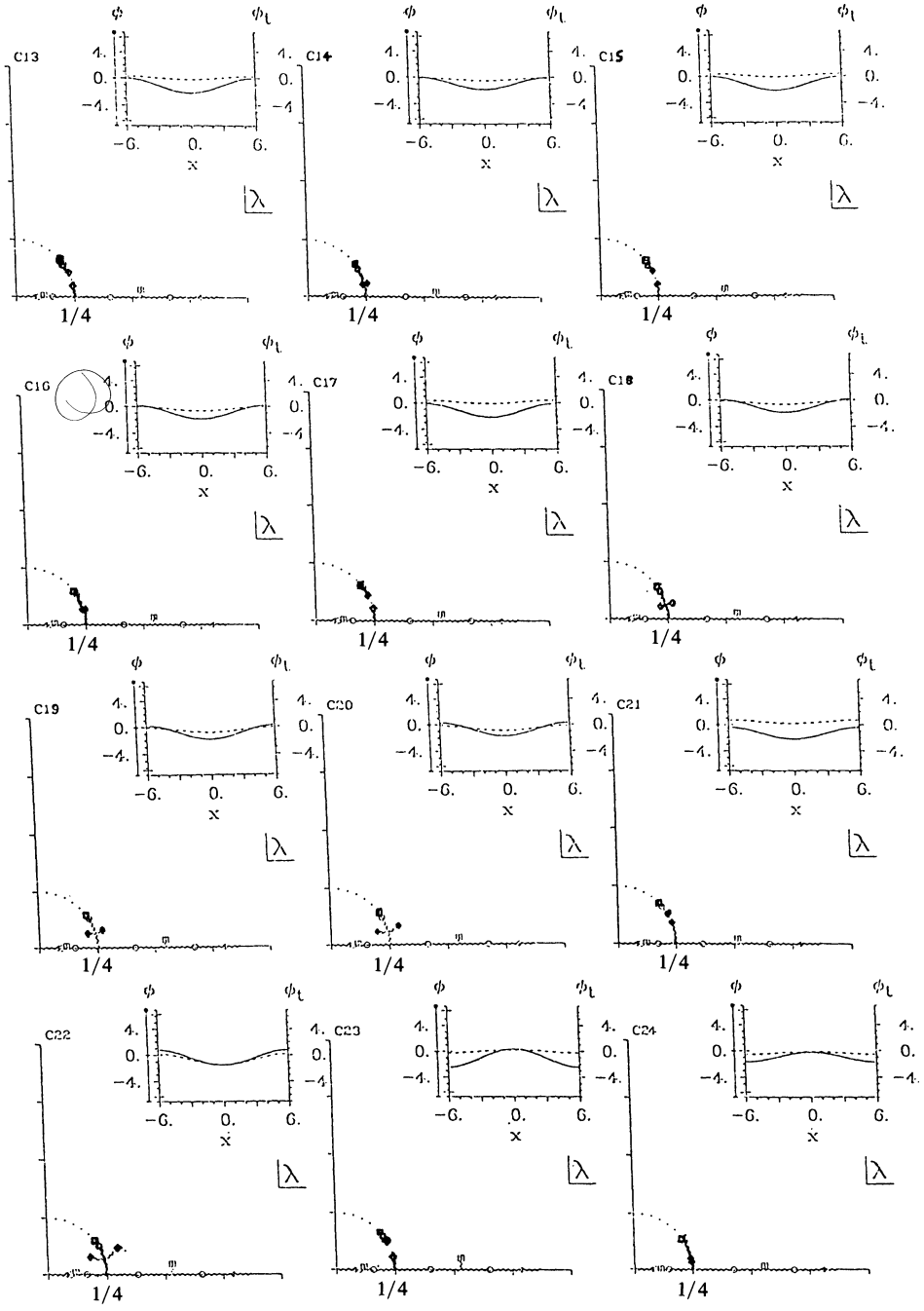


FIG. 14—continued

The agreement is quite good. When the odes pass from $h>0$ to $h<0$, the sine-Gordon projection goes from a "gap" to a "cross" configuration. The agreement is not one-to-one when the odes are very close to the exact homoclinic structure ($h\approx 0$), but this is expected due to the approximation by linear Fourier modes.

6. Concluding remarks. Based on the combination of:

(i) Our geometric understanding of the integrable sine-Gordon phase space, with singular components homoclinic to tori, and

(ii) The numerically verified presence of these singular components in the chaotic dynamics of the weakly perturbed system,

we are led to a research program to coordinatize these chaotic attractors and model the dynamics with associated amplitude equations. This paper represents the simplest example in the small amplitude limit of this general program. This two complex amplitude truncation has already yielded excellent correlation with the homoclinic structure that it shares with the perturbed integrable pde. Moreover, this "physically derived" four-dimensional dynamical system is a fertile example from the pure dynamical systems point of view. A Melnikov-type analysis based on the homoclinic orbits has now been developed for this model problem [4a], [4b], and will appear in the thesis of Kovacic.

The next step in this program is to truncate on the fully nonlinear sine-Gordon or nonlinear Schrödinger modes. Thus far, we have derived the averaged finite amplitude modal equations for systems (1.1) and (2.2), and have explicit formulas for the $K_0\oplus K_1$ truncation relevant for the study presented here. (This analysis is discussed in [7].) The averaged equations are equivalent to the order ε rate of change for each sine-Gordon integral in the presence of the perturbation, i.e., $dH_j/dt=\varepsilon f_j$, $j\in Z$. These are precisely the quantities required in a Melnikov analysis in higher dimensions.

To capture the dynamics of the chaotic attractors, we must couple the rapid phases (the angles) to the averaged equations (the actions) described above. These perturbed action-angle equations will be reported in [5]. The next nontrivial step is to numerically analyze these truncated dynamical systems. Since these nonlinear coordinates are naturally defined on Riemann surfaces (see [3], [5]–[7]), we are currently developing efficient algorithms for these computations [8].

Appendix. Most of the numerical runs were done using a second-order in time, fourth-order in space discretization of the sine-Gordon pde (1.1). The u_{xx} term was discretized by fourth-order central differences with $\Delta x=0.20$ and the u_t and u_{tt} terms were discretized by second-order central differences with $\Delta t=0.02$ (i.e., this is a leapfrog scheme). The initial timestep was calculated by Taylor series, namely,

$$u(x,\, t=\Delta t)=u(x,0)+u_t(x,0)\Delta t+\tfrac{1}{2}u_{tt}(x,0)(\Delta t)^2$$

and the u_{tt} term was replaced by using (1.1a). Selected runs were rechecked by a fourth-order Runge–Kutta method in time using eighth-order central differences in space (and the same Δx and Δt as above) on the pde to make sure that the long-time evolution of the run was as indicated by the lower-order method. In addition the boundaries between periodic and chaotic runs in time were determined to three decimal places by this higher-order code.

Acknowledgments. We express gratitude to Mac Hyman, Michael Jolly, and Yannis Kevrekides for generous numerical assistance. M. Gregory Forest and David W. McLaughlin also thank L. S. Young for helpful discussions. We all express gratitude to the Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratories, Los Alamos, New Mexico, where this research was performed.

REFERENCES

[1] A. R. BISHOP, M. G. FOREST, D. W. MCLAUGHLIN, AND E. A. OVERMAN II, *A quasi-periodic route to chaos in a near-integrable pde*, Phys. D, 23 (1986), pp. 293–328, and references cited therein.

[2] A. R. BISHOP, D. W. MCLAUGHLIN, AND E. A. OVERMAN II, *Coherence and chaos in the driven, damped sine-Gordon equation: measurement of the soliton spectrum*, Phys. D, 19 (1986), pp. 1–41.

[3a] N. M. ERCOLANI, M. G. FOREST, AND D. W. MCLAUGHLIN, *Geometry of the Modulational Instability. Part* I: *Local Analysis*, Mem. Amer. Math. Soc., to appear.

[3b] ———, *Geometry of the Modulational Instability. Part* II: *Global Analysis*, Mem. Amer. Math. Soc., to appear.

[3c] ———, *Homoclinic orbits for the periodic sine-Gordon equation*, Phys. D, to appear.

[3d] ———, *The origin and saturation of modulational instabilities*, Phys. D, 18 (1986), pp. 472–474.

[4a] G. KOVACIC AND S. WIGGINS, preprint, California Institute of Technology, Pasadena, CA, 1988. Phys. D, submitted.

[4b] N. M. ERCOLANI, M. G. FOREST, AND D. W. MCLAUGHLIN, *Notes on Melnikov integrals for models of the driven pendulum chain*, preprint, University of Arizona, Tempe, AZ, 1988.

[5] ———, *Fully nonlinear modal equations for nearly integrable pdes*, preprint, Ohio State University, Columbus, OH, May 1988.

[6a] N. M. ERCOLANI, M. G. FOREST, D. W. MCLAUGHLIN, AND R. MONTGOMERY, *Hamiltonian structure of the modulation equations for sine-Gordon wavetrains*, Duke Math. J., 55 (1987), pp. 949–983.

[6b] M. G. FOREST AND D. W. MCLAUGHLIN, *Canonical variables of the periodic sine-Gordon equation and a method of averaging*, unpublished preprint, internal report of the Los Alamos National Laboratory, Los Alamos, NM, 1978.

[7] N. M. ERCOLANI, M. G. FOREST, AND D. W. MCLAUGHLIN, *Oscillations and instabilities in nearly integrable pdes*, Lectures in Appl. Math., 23 (1985), pp. 3–46.

[8] R. FLESCH, M. G. FOREST, AND A. SINHA, *Numerical inverse spectral transform for the periodic sine-Gordon equation: theta function solutions and their linearized stability*, Phys. D, submitted.

[9] E. OTT AND D. A. RUSSELL, *Chaotic (strange) and periodic behavior in instability saturation by the oscillating two-stream instability*, Phys. Fluids, 24 (1981), pp. 1976–1988.

[10] J. C. ARIYASU AND A. R. BISHOP, Phys. Rev. B, (3), (1987).

[11] A. R. BISHOP, M. G. FOREST, D. W. MCLAUGHLIN, AND E. A. OVERMAN II, *Quasiperiodic route to chaos in a near integrable P.D.E.: homoclinic crossings*, Phys. Lett. A, 127 (1988), pp. 335–340.

[12] D. W. MCLAUGHLIN, A. PEARLSTEIN, AND G. TERRONES, *Stability and bifurcation of time-periodic solutions of the damped and driven sine-Gordon equation*, preprint, University of Arizona, Tempe, AZ, May 1988.

[13] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Appl. Math. Sci., Vol. 42, Springer-Verlag, Berlin, New York, 1983.

[14] S. WIGGINS, *Global Bifurcations and Chaos: Analytical Methods*, Appl. Math. Sci., Vol. 73, Springer-Verlag, Berlin, New York, 1988.

[15] N. M. ERCOLANI, M. G. FOREST, G. KOVACIC, D. W. MCLAUGHLIN, AND S. WIGGINS, work in progress.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1537-1549, November 1990
© 1990 Society for Industrial and Applied Mathematics
009

FAR-FIELD PATTERNS FOR ELECTROMAGNETIC WAVES IN AN INHOMOGENEOUS MEDIUM*

DAVID COLTON† AND LASSI PÄIVÄRINTA‡

Abstract. This paper considers the scattering of time-harmonic electromagnetic waves by an inhomogeneous medium of compact support, i.e., the permittivity $\varepsilon = \varepsilon(x)$ and the conductivity $\sigma = \sigma(x)$ are functions of $x \in R^3$. If $\sigma > 0$, it is shown that the set of far-field patterns of the electric fields corresponding to incident plane waves propagating in arbitrary directions with arbitrary polarization is complete in the space of square integrable tangential vector fields defined on the unit sphere. On the other hand, if $\sigma = 0$ it is shown that for the case of a spherically stratified medium there exist values of the frequency such that the set of far-field patterns is not complete. Finally, it is shown that, if from each far-field pattern is subtracted the electric far-field pattern corresponding to an electromagnetic field satisfying an impedance boundary condition on the boundary of a ball containing the inhomogeneity, then the resulting class is complete for $\sigma \geqq 0$ and $\varepsilon \geqq 0$.

Key words. far-field patterns, electromagnetic waves

AMS(MOS) subject classifications. 35P25, 78A45

1. Introduction. In this paper, we are interested in the class $\mathcal{F}$ of electric far-field patterns corresponding to the scattering of time-harmonic electromagnetic plane waves by an inhomogeneous medium of compact support. The corresponding problem for acoustic waves has been considered in [2]. However, for electromagnetic waves the class $\mathcal{F}$ has only been considered for the case of constant permittivity and variable conductivity, with the conclusion that $\mathcal{F}$ is complete in the space of square integrable tangential vector fields defined on the unit sphere [8]. Results of this nature are of particular interest in inverse scattering theory since they form the theoretical basis for new algorithms for solving the inverse scattering problem [6]-[8].

Continuing the work initiated in [8], we will now concern ourselves with the class $\mathcal{F}$ of electric far-field patterns in the case when both the permittivity and conductivity are functions of $x \in R^3$. This is the case, for example, in many applications in medical imaging [14]. We begin by deriving a reciprocity relation and using it to show that $\mathcal{F}$ is complete if and only if the solution of the interior transmission problem (first introduced in [8]) is unique. We then show that this is the case if the conductivity is positive (in the special case of constant permittivity, this provides a new proof for Theorem 2.1 of [8]). However, for the case of vanishing conductivity and spherically stratified permittivities we will show that there always exists a discrete set of frequencies such that $\mathcal{F}$ is not complete. This analysis is based on an asymptotic analysis of the radial Schrödinger equation as discussed in § 12.6 of [13]. Finally, we show that if from each element in $\mathcal{F}$ we subtract the electric far-field pattern corresponding to an electromagnetic field satisfying an impedance boundary condition on the boundary of a ball containing the inhomogeneous medium in its interior, then the resulting class of functions is complete for all conductivities $\sigma \geqq 0$ and permittivities $\varepsilon \geqq 0$. The proof of this result again relies heavily on the reciprocity principle.

* Received by the editors September 18, 1989; accepted for publication (in revised form) December 1, 1989.

† Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19716. The research of this author was supported in part by grants from the Air Force Office of Scientific Research and the National Science Foundation.

‡ Department of Mathematics, University of Helsinki, Helsinki, Finland.

In what follows, we denote the vector product of two vectors by $[\cdot, \cdot]$ and the triple product of three vectors by $(\cdot, \cdot, \cdot)$. A unit vector in R^3 will always be designated by a caret, i.e., $\hat{x}$, $\hat{\alpha}$, etc.

2. Reciprocity and the interior transmission problem for electromagnetic waves. Consider the propagation of a time-harmonic electromagnetic wave through a medium of variable conductivity and permittivity. Let $\sigma(x)$ denote the conductivity and $\varepsilon(x)$ the permittivity for $x \in R^3$. Assume that $\sigma(x) \geqq 0$ for all x and $\sigma(x) = 0$ for $|x| \geqq a > 0$, whereas $\varepsilon(x) \geqq 0$ for all x and $\varepsilon(x) = \varepsilon_0$ for $|x| \geqq a$, where ε_0 is the dielectric constant. It is assumed that a is chosen such that the supports of $\sigma(x)$ and $\varepsilon(x) - \varepsilon_0$ are properly contained in

$$B = \{x: |x| < a\}.$$

Then if ω is the frequency of the incident field and μ_0 is the permeability, the direct scattering problem is to determine the electric field $E(x)$ and the magnetic field $H(x)$ such that Maxwell's equations

$$\begin{aligned} &\operatorname{curl} E - i\omega\mu_0 H = 0, \\ &\operatorname{curl} H + i\omega\varepsilon(x)E = \sigma(x)E \end{aligned} \tag{2.1}$$

are satisfied for $x \in R^3$, where $E(x) = E^i(x) + E^s(x)$, $H(x) = H^i(x) + H^s(x)$ with $\{E^i, H^i\}$ being the incident field and $\{E^s, H^s\}$ the scattered field. In particular, $\{E^s, H^s\}$ is required to satisfy the Silver-Müller radiation condition [3], [12]. We will assume that $\sigma(x)$ and $\varepsilon(x)$ are in Hölder class $C^{2,\alpha}$, $0 < \alpha < 1$, thus allowing us to pass freely between (2.1) and the vector Helmholtz equation [5].

Now consider the set of incident fields of the form

$$\begin{aligned} H^i(x; \hat{\alpha}, p) &= \operatorname{curl} p \exp[ikx \cdot \hat{\alpha}], \\ E^i(x; \hat{\alpha}, p) &= -\frac{1}{i\varepsilon_0\omega} \operatorname{curl} H^i(x; \hat{\alpha}, p) \end{aligned} \tag{2.2}$$

where $x \in R^3$, $\hat{\alpha}$ is a unit vector giving the direction of propagation, $k^2 = \varepsilon_0\mu_0\omega^2$ and p is a constant vector giving the polarization. From Corollary 4.9 of [3] we see that

$$\begin{aligned} E^s(x; \hat{\alpha}, p) &= \frac{e^{ik|x|}}{|x|} F(\hat{x}; \hat{\alpha}, p) + O\left(\frac{1}{|x|^2}\right), \\ H^s(x; \hat{\alpha}, p) &= \frac{e^{ik|x|}}{|x|} [\hat{x}, F(\hat{x}; \hat{\alpha}, p)] + O\left(\frac{1}{|x|^2}\right) \end{aligned} \tag{2.3}$$

where $\hat{x} = x/|x|$ and F is the electric far-field pattern. It has been shown in [1] that if $q \in R^3$, then $q \cdot F$ has the representation

$$\begin{aligned} q \cdot F(\hat{x}; \hat{\alpha}, p) = \frac{i}{\mu_0\omega} \int_{\partial B} \{&(\hat{y}, \operatorname{curl} H^s(y; \hat{\alpha}, p), H^i(y; -\hat{x}, q)) \\ &+ (\hat{y}, H^s(y; \hat{\alpha}, p), \operatorname{curl} H^i(y; -\hat{x}, q))\}\, ds\,(y). \end{aligned} \tag{2.4}$$

Note that since

$$\begin{aligned} \operatorname{curl}_y qe^{-ik\hat{x}\cdot y} &= ik[q, \hat{x}]e^{-ik\hat{x}\cdot y}, \\ \operatorname{curl}_y \operatorname{curl}_y q\bar{e}^{ik\hat{x}\cdot y} &= k^2(q - (q \cdot \hat{x})\hat{x})\, e^{-ik\hat{x}\cdot y}, \end{aligned} \tag{2.5}$$

we have that if $q = \hat{x}$ then $\hat{x} \cdot F = 0$, i.e., the electric far-field pattern is tangential to the unit sphere.

We will now prove the following reciprocity theorem for the electric far-field patterns corresponding to the scattering problem (2.1), (2.2). Versions of the following theorem are well known (cf. [11, p. 64]); however, we will need the precise form given below and hence prove it here. In what follows, $\partial\Omega$ denotes the unit sphere in R^3.

THEOREM 2.1. *For all vectors $\hat{\alpha}, \hat{x} \in \partial\Omega$ and $p, q \in R^3$ we have*

$$q \cdot F(\hat{x}; \hat{\alpha}, p) = p \cdot F(-\hat{\alpha}, -\hat{x}, q).$$

Proof. We begin by establishing three integral identities we need to prove the theorem. Let $F_h = [\hat{x}, F]$. Then

(2.6)

$$\int_{\partial B} \{(\hat{y}, \operatorname{curl} H^s(y; \hat{\alpha}, p), H^s(y; -\hat{x}, q)) + (\hat{y}, H^s(y; \hat{\alpha}, p), \operatorname{curl} H^s(y; -\hat{x}, q))\}\, ds$$

$$= -i\varepsilon_0\omega \int_{\partial B} \{(\hat{y}, E^s(y; \hat{\alpha}, p), H^s(y; -\hat{x}, q)) + (\hat{y}, H^s(y; \hat{\alpha}, p), E^s(y; -\hat{x}, q))\}\, ds$$

$$= -i\varepsilon_0\omega\, e^{2ika} \int_{\partial\Omega} \{(\hat{y}, F(\hat{y}; \hat{\alpha}, p), F_h(\hat{y}; -\hat{x}, q)) + (\hat{y}, F_h(\hat{y}; \hat{\alpha}, p), F(\hat{y}; -\hat{x}, q))\}\, ds$$

$$+ O\left(\frac{1}{a}\right).$$

Since

$$\begin{aligned}(2.7)\qquad [F(\hat{y}; \hat{\alpha}, p), F_h(\hat{y}; -\hat{x}, q)] &= [F(\hat{y}; \hat{\alpha}, p), [\hat{y}, F(\hat{y}; -\hat{x}, q)]] \\ &= F(\hat{y}; -\hat{x}, q) \cdot F(\hat{y}; \hat{\alpha}, p)\hat{y},\end{aligned}$$

$$\begin{aligned}(2.8)\qquad [F_h(\hat{y}; \hat{\alpha}, p), F(y; -\hat{x}, q)] &= [[\hat{y}, F(\hat{y}; \hat{\alpha}, p)], F(\hat{y}; -\hat{x}, q)] \\ &= -F(\hat{y}; -\hat{x}, q) \cdot F(\hat{y}; \hat{\alpha}, p)\hat{y},\end{aligned}$$

we have, letting $a \to \infty$, that

(2.9)

$$\int_B \{(\hat{y}, \operatorname{curl} H^s(y; \hat{\alpha}, p), H^s(y; -\hat{x}, q)) + (\hat{y}, H^s(y; \hat{\alpha}, p), \operatorname{curl} H^s(y; -\hat{x}, q))\}\, ds = 0.$$

This is the first identity we need.

To derive our second identity, we use the second vector Green's theorem ([3, p. 117]) and the fact that, from (2.1) and the regularity assumptions on $\sigma(x)$ and $\varepsilon(x)$,

$$\begin{aligned}(2.10)\qquad &\Delta E - \operatorname{grad} \operatorname{div} E + i\omega\mu_0\sigma(x)E + \omega^2\mu_0\varepsilon(x)E = 0, \qquad x \in R^3, \\ &\operatorname{div} E(x) = 0, \qquad |x| \geqq a\end{aligned}$$

to deduce that

$$\begin{aligned}
&\int_{\partial B} \{(\hat{y}, \operatorname{curl} H(y; \hat{\alpha}, p), H(y; -\hat{x}, q)) + (\hat{y}, H(y; \hat{\alpha}, p), \operatorname{curl} H(y; -\hat{x}, q))\, ds \\
&\quad = -\frac{\varepsilon_0}{\mu_0} \int_{\partial B} \{(\hat{y}, E(y; \hat{\alpha}, p), \operatorname{curl} E(y, -\hat{x}, q)) \\
&\qquad\qquad + (\hat{y}, \operatorname{curl} E(y; \hat{\alpha}, p), E(y; -\hat{x}, q))\}\, ds \\
&\quad = \int_B \int \{E(y; \hat{\alpha}, p) \cdot \Delta E(y; -\hat{x}, q) - E(y; -\hat{x}, q) \cdot \Delta E(y; \hat{\alpha}, p)\}\, dx \qquad (2.11) \\
&\quad = \int_B \int \{E(y; \hat{\alpha}, p) \cdot \operatorname{grad} \operatorname{div} E(y; -\hat{x}, q) \\
&\qquad\qquad - E(y; -\hat{x}, q) \cdot \operatorname{grad} \operatorname{div} E(y; \hat{\alpha}, p)\}\, dx \\
&\quad = -\int_B \int \{\operatorname{div} E(y; \hat{\alpha}, p) \operatorname{div} E(y; -\hat{x}, q) \\
&\qquad\qquad - \operatorname{div} E(y; -\hat{x}, q) \operatorname{div} E(y; \hat{\alpha}, p)\}\, dx = 0.
\end{aligned}$$

This is the second identity we need.

The final identity we shall need can be derived in exactly the same manner as (2.11) and we simply state the result:

(2.12)

$$\int_{\partial B} \{(\hat{y}, \operatorname{curl} H^i(y; \hat{\alpha}, p), H^i(y; -\hat{x}, q)) + (\hat{y}, H^i(y; \hat{\alpha}, p), \operatorname{curl} H^i(y; -\hat{x}, q))\}\, ds = 0.$$

We can now prove our theorem. Using (2.4) and the above identities, we have

$$\begin{aligned}
q \cdot F(\hat{x}; \hat{\alpha}, p) &= \frac{i}{\mu_0 \omega} \int_{\partial B} \{(\hat{y}, \operatorname{curl} H^s(y; \hat{\alpha}, p), H^i(y; -\hat{x}, q)) \\
&\qquad + (\hat{y}, H^s(y; \hat{\alpha}, p), \operatorname{curl} H^i(y; -\hat{x}, q))\}\, ds \\
&= \frac{i}{\mu_0 \omega} \int_{\partial B} \{(\hat{y}, \operatorname{curl} H^s(y; \hat{\alpha}, p), (H(y; -\hat{x}, q) - H^s(y; -\hat{x}, q))) \\
&\qquad + (\hat{y}, H^s(y; \hat{\alpha}, p), \operatorname{curl} (H(y; -\hat{x}, q) - H^s(y; -\hat{x}, q)))\}\, ds \\
&= \frac{i}{\mu_0 \omega} \int_{\partial B} \{(\hat{y}, \operatorname{curl} (H(y; \hat{\alpha}, p) - H^i(y; \hat{\alpha}, p)), H(y; -\hat{x}, q)) \\
&\qquad + (\hat{y}, (H(y; \hat{\alpha}, p) - H^i(y; \hat{\alpha}, p)), \operatorname{curl} H(y; -\hat{x}, q))\}\, ds \qquad (2.13) \\
&= -\frac{i}{\mu_0 \omega} \int_{\partial B} \{(\hat{y}, \operatorname{curl} H^i(y; \hat{\alpha}, p), H(y; -\hat{x}, q)) \\
&\qquad + (\hat{y}, H^i(y; \hat{\alpha}, p), \operatorname{curl} H(y; -\hat{x}, q))\}\, ds \\
&= \frac{i}{\mu_0 \omega} \int_{\partial B} \{(\hat{y}, \operatorname{curl} H^s(y; -\hat{x}, q), H^i(y; \hat{\alpha}, p)) \\
&\qquad + (\hat{y}, H^s(y; -\hat{x}, q), \operatorname{curl} H^i(y; \hat{\alpha}, p))\}\, ds \\
&= p \cdot F(-\hat{\alpha}, -\hat{x}, q).
\end{aligned}$$

This completes the proof of the theorem.

We now want to use the above reciprocity theorem to show the equivalence of the completeness of the set of electric far-field patterns and the uniqueness of the solution to the *interior transmission problem*: Find E_0, E_1, H_0, $H_1 \in C^1(B) \cap C(\bar{B})$ such that

$$\begin{aligned} &\operatorname{curl} E_1 - i\omega\mu_0 H_1 = 0 \\ &\operatorname{curl} H_1 + i\omega\varepsilon(x) E_1 = \sigma(x) E_1 \end{aligned} \quad \text{in } B, \tag{2.14}$$

$$\begin{aligned} &\operatorname{curl} E_0 - i\omega\mu_0 H_0 = 0 \\ &\operatorname{curl} H_0 + i\omega\varepsilon_0 E_0 = 0 \end{aligned} \quad \text{in } B, \tag{2.15}$$

$$\begin{aligned} &[\hat{x}, E_1 - E_0] = 0 \\ &[\hat{x}, H_1 - H_0] = 0 \end{aligned} \quad \text{on } \partial B \tag{2.16}$$

where again $B = \{x : |x| < a\}$. To this end, define the Hilbert space Y by

$$Y = \{g : g \in L^2(\partial\Omega),\ \hat{x} \cdot g(\hat{x}) = 0\},$$

let $\{\hat{\alpha}_n\}_{n=1}^{\infty}$ be a dense set of unit vectors on $\partial\Omega$ (i.e., every surface patch on $\partial\Omega$ contains at least one vector of this set), and consider the set $\mathcal{F}$ of electric far-field patterns defined by

$$\mathcal{F} = \{F(\hat{x}; \hat{\alpha}_n, \hat{e}_l) : l = 1, 2, 3,\ n = 1, 2, 3, \cdots\}$$

where $\{\hat{e}_l\}_{l=1}^{3}$ are the unit coordinate vectors in R^3. A solution $\{E, H\}$ of Maxwell's equations (2.1) with $\varepsilon(x) = \varepsilon_0$ and $\sigma(x) = 0$ will be called an *electromagnetic Herglotz pair* if there exists $g \in Y$ such that

$$E(x) = \int_{\partial\Omega} g(\hat{y}) \exp[ikx \cdot \hat{y}]\, ds\,(\hat{y}) \tag{2.17}$$

where again $k^2 = \varepsilon_0\mu_0\omega^2$. The function g is called the *Herglotz kernel* of E, and it is easily seen from Sonine's formula for Bessel functions that $g = 0$ if and only if E is identically zero. We can now prove the following theorem, where $\mathcal{F}^{\perp}$ denotes the orthogonal complement of $\mathcal{F}$ in Y.

THEOREM 2.2. *$g \in \mathcal{F}^{\perp}$ if and only if there exists a solution of the interior transmission problem such that $\{E_0, H_0\}$ is an electromagnetic Herglotz pair with Herglotz kernel g.*

Proof. Suppose there exists a $g \in Y$ such that

$$\int_{\partial\Omega} F(\hat{x}; \hat{\alpha}_n, \hat{e}_l) \cdot \overline{g(\hat{x})}\, ds\,(\hat{x}) = 0 \tag{2.18}$$

for $l = 1, 2, 3$, $n = 1, 2, 3, \cdots$. Then by continuity and superposition we have that

$$\int_{\partial\Omega} F(\hat{x}; \hat{\alpha}, p) \cdot \overline{g(\hat{x})}\, ds\,(\hat{x}) = 0 \tag{2.19}$$

for all $\hat{\alpha} \in \partial\Omega$, $p \in R^3$. By Theorem 2.1 this is equivalent to

$$p \cdot \int_{\partial\Omega} F(-\hat{\alpha}; -\hat{x}, \overline{g(\hat{x})})\, ds\,(\hat{x}) = 0 \tag{2.20}$$

for all $\hat{\alpha} \in \partial\Omega$, $p \in R^3$, i.e.,

$$\int_{\partial\Omega} F(\hat{x}; \hat{\alpha}, h(\hat{\alpha}))\, ds\,(\hat{\alpha}) = 0 \tag{2.21}$$

for all $\hat{x} \in \partial\Omega$, where $h(\hat{\alpha}) = \overline{g(-\hat{\alpha})}$.

Now define the electromagnetic Herglotz pair $\{E_0, H_0\}$ by

$$\begin{aligned} H_0(x) &= \int_{\partial\Omega} H^i(x; \hat{\alpha}, h(\hat{\alpha}))\, ds(\hat{\alpha}) \\ &= \operatorname{curl} \int_{\partial\Omega} h(\hat{\alpha}) \exp[ikx \cdot \hat{\alpha}]\, ds\,(\hat{\alpha}), \\ E_0(x) &= -\frac{1}{i\omega\varepsilon_0} \operatorname{curl} H_0(x) \\ &= \mu_0\omega \int_{\partial\Omega} h(\hat{\alpha}) \exp[ikx \cdot \hat{\alpha}]\, ds\,(\hat{\alpha}). \end{aligned} \tag{2.22}$$

Then, from (2.1), (2.2), and (2.21), the scattered field $\{E_0^s, H_0^s\}$ corresponding to the incident field $\{E_0, H_0\}$ vanishes for $|x| \geqq a$, i.e., $E_0(x) = E_1(x)$, $H_0(x) = H_1(x)$ for $|x| \geqq a$, where $E_1(x) = E_0(x) + E_0^s(x)$, $H_1(x) = H_0(x) + H_0^s(x)$. But $\{E_1, H_1\}$ satisfies (2.1) for $x \in R^3$. Hence, E_0, E_1, H_0, H_1 satisfies the interior transmission problem (2.14)-(2.16). Conversely, if there exists a solution to (2.14)-(2.16) such that $\{E_0, H_0\}$ is an electromagnetic Herglotz pair, then, using the representation theorem for Maxwell's equations and unique continuation, we can show (cf. Theorem 2.3 below) that $E_1(x)$ and $H_1(x)$ are defined for all $x \in R^3$, $E_1(x) = E_0(x)$, and $H_1(x) = H_0(x)$ for $|x| \geqq a$, and $\{E_0^s(x), H_0^s(x)\}$, as defined above, satisfies the radiation condition. By uniqueness of the solution to (2.1), (2.2) [12] we now have that (2.21) is valid. This completes the proof of the theorem.

Showing that two statements are equivalent is useful only if one of the statements is easier to apply than the other. As an example of the usefulness of Theorem 2.2, we have Theorem 2.3 below. Recall that if D denotes the support of $\sigma(x)$ then D is properly contained in $B = \{x : |x| < a\}$. We assume without loss of generality that ∂D is smooth and denote the unit outward normal to ∂D by $\hat{n}$.

THEOREM 2.3. *Suppose D is nonempty and the support of $\varepsilon(x) - \varepsilon_0$ is contained in D. Then $\mathcal{F}$ is complete in Y.*

Proof. By Theorem 2.2, it suffices to show that the only solution of the interior transmission problem (2.14)-(2.16) is E_0, E_1, H_0, H_1 identically zero. Applying the representation theorem for Maxwell's equations [3, p. 110] to $B \setminus \bar{D}$ shows that $E_1 - E_0$ can be continued into $R^3 \setminus \bar{D}$ and satisfies the radiation condition. By the uniqueness of the exterior Maxwell boundary value problem [3, p. 126], we see that $E_1 = E_0$ in $R^3 \setminus B$ and, by unique continuation, in $R^3 \setminus \bar{D}$. Using the regularity assumptions on $\sigma(x)$ and $\varepsilon(x)$, we can now replace (2.14)-(2.16) by the equivalent system

$$\begin{aligned} &\Delta E_1 - \operatorname{grad} \operatorname{div} E_1 + \omega^2\mu_0\varepsilon(x)E_1 + i\omega\mu_0\sigma(x)E_1 = 0 \\ &\Delta E_0 + \omega^2\mu_0\varepsilon_0 E_0 = 0 \end{aligned} \quad \text{in } B, \tag{2.23}$$

$$E_1 = E_0 \quad \text{in } \bar{B} \setminus D. \tag{2.24}$$

Noting that $\operatorname{div} E_0 = 0$ in D, we now apply the first vector Green's theorem [3, p. 117]

to deduce that

$$
\begin{aligned}
0 &= \int_{\partial D} \{(\hat{n}, \bar{E}_0, \operatorname{curl} E_0) - (\hat{n}, \bar{E}_1, \operatorname{curl} E_1)\}\, ds \\
&= \int_D \int (|\operatorname{curl} E_0|^2 - \omega^2 \varepsilon_0 \mu_0 |E_0|^2)\, dx - \int_{\partial D} (\hat{n}, \bar{E}_1, \operatorname{curl} E_1)\, ds,
\end{aligned} \tag{2.25}
$$

i.e.,

$$
\operatorname{Im} \int_{\partial D} (\hat{n}, \bar{E}_1, \operatorname{curl} E_1)\, ds = 0. \tag{2.26}
$$

But the first vector Green's theorem gives

$$
\int_D \int (\bar{E}_1 \cdot \Delta E_1 + |\operatorname{curl} E_1|^2 + |\operatorname{div} E_1|^2)\, dx = \int_{\partial D} (\hat{n}, \bar{E}_1, \operatorname{curl} E_1)\, ds, \tag{2.27}
$$

i.e.,

$$
\operatorname{Im} \int_D \int \bar{E}_1 \cdot \Delta E_1\, dx = 0. \tag{2.28}
$$

Using the fact that $\operatorname{div} E_1 = 0$ on ∂D, we have

$$
\begin{aligned}
\int_D \int \bar{E}_1 \cdot \Delta E_1\, dx &= \int_D \int \bar{E}_1 \cdot (\operatorname{grad} \operatorname{div} E_1 - \omega^2 \mu_0 \varepsilon(x) E_1 - i\omega \mu_0 \sigma(x) E_1)\, dx \\
&= -\int_D \int (|\operatorname{div} E_1|^2 + \omega^2 \mu_0 \varepsilon(x) |E_1|^2 + i\omega \mu_0 \sigma(x) |E_1|^2)\, dx.
\end{aligned} \tag{2.29}
$$

From (2.28), (2.29) we now have that

$$
\int_D \int \sigma(x) |E_1(x)|^2\, dx = 0, \tag{2.30}
$$

i.e., $E(x) = 0$ for $x \in D$.

Since $E_1(x) = 0$ for $x \in D$ and $E_1 \in C^1(B)$, we have that $E_1(x)$ and $\operatorname{curl} E_1(x)$ vanish for $x \in \partial D$. From the representation theorem for solutions of Maxwell's equations defined in $B \setminus \bar{D}$ [3, p. 110], we can now conclude that $E_1(x) = 0$ for $x \in B$ and hence that $H_1(x) = 0$ for $x \in B$ also. From (2.16) and the representation theorem again, we have that $E_0(x) = 0$ for $x \in D$ and hence that $H_0(x) = 0$ for $x \in D$ also. The theorem is now proved.

The situation in which $\sigma(x)$ is identically zero is more complicated. We will examine this case for spherically stratified inhomogeneities in the next section of this paper.

3. Nonconducting spherically stratified media. In this section of our paper, we again want to apply Theorem 2.2, but now for the situation in which $\sigma(x)$ is identically zero and $\varepsilon(x) = \varepsilon(r)$, $r = |x|$, is spherically stratified. As will be seen, in this case there exist frequencies ω such that $\mathscr{F}$ is not complete. To show this, we see from Theorem 2.2 that we need to examine when there exist nontrivial solutions to the interior transmission problem (2.14)-(2.16).

Assume $\sigma(x)=0$ and $\varepsilon(x)=\varepsilon(r)$, where $\varepsilon(r)>0$ for $0\leqq r<a$ and $\varepsilon(r)=\varepsilon_0$ for $r\geqq a$. Following [10], we introduce the Hertz vectors $C=(ru(r,\theta),0,0)$ and $C_0=(rh(r,\theta),0,0)$, where (r,θ,ϕ) are spherical coordinates and u and h are scalar functions. Then (cf. [10])

$$
\begin{aligned}
E_1(x)&=\frac{i\omega\mu_0}{\varepsilon(r)}\operatorname{curl}\varepsilon(r)C,\\
H_1(x)&=\omega^2\mu_0\varepsilon(r)C+\operatorname{grad}\frac{\partial}{\partial r}(ru),\\
E_0(x)&=i\omega\mu_0\operatorname{curl}C_0,\\
H_0(x)&=\omega^2\mu_0\varepsilon_0C_0+\operatorname{grad}\frac{\partial}{\partial r}(rh)
\end{aligned}
\tag{3.1}
$$

will be a solution of the interior transmission problem (2.14)-(2.16) provided $\{u,h\}$ is a solution of the interior acoustic transmission problem

$$
\begin{aligned}
&\Delta u+k^2n(r)u=0\\
&\Delta h+k^2h=0
\end{aligned}
\qquad \text{in } B, \tag{3.2}
$$

$$
u=h \tag{3.3}
$$

$$
\frac{\partial u}{\partial r}=\frac{\partial h}{\partial r} \qquad \text{on } \partial B
$$

where $k^2=\omega^2\mu_0\varepsilon_0$, $n(r)=\varepsilon(r)/\varepsilon_0$.

To find a nontrivial solution of (3.2), (3.3), we expand u and h in a Legendre series

$$
\begin{aligned}
u(r,\theta)&=\sum_{l=0}^{\infty}a_l\frac{1}{r}v_l(r)P_l(\cos\theta),\\
h(r,\theta)&=\sum_{l=0}^{\infty}b_lj_l(kr)P_l(\cos\theta)
\end{aligned}
\tag{3.4}
$$

where P_l is Legendre's polynomial, j_l is a spherical Bessel function, a_l and b_l are constants to be determined, and

$$
\ddot{v}_l+\left(k^2n(r)-\frac{l(l+1)}{r^2}\right)v_l=0 \tag{3.5}
$$

for $0\leqq r\leqq a$, where $v_l(r)\sim rj_l(kr)$ as $n\to 1$ in the maximum norm and $v_l\in C[0,a]$. Using Green's function for Bessel's equation, it can be shown that under these assumptions v_l is uniquely determined. From (3.1) it can be seen that the $l=0$ term in (3.4) yields a solution of the interior transmission problem (2.14)-(2.16) that is identically zero. Hence, we are only interested in the components

$$
ru_l(r,\theta)=a_lv_l(r)P_l(\cos\theta), \qquad h_l(r,\theta)=b_lj_l(kr)P_l(\cos\theta) \tag{3.6}
$$

for $l\geqq 1$. We will show that if $\varepsilon(r)>\varepsilon_0$ for $0\leqq r<a$ or $\varepsilon(r)<\varepsilon_0$ for $0\leqq r<a$, then for each $l\geqq 1$ there exists an infinite set of values of k and constants $a_l=a_l(k)$, $b_l=b_l(k)$, such that (3.6) is a nontrivial solution of (3.2), (3.3), which by (3.1) defines a nontrivial solution of (2.14)-(2.16). From (3.1) and (3.6), it is easily seen that $\{E_0,H_0\}$ is an electromagnetic Herglotz pair. Hence, by Theorem 2.2, for such values of k the set of electric far-field patterns is not complete.

To show the existence of values of k such that (3.6) is a nontrivial solution of (3.2), (3.3), we need to examine the asymptotic behavior of the solution of (3.5) satisfying the initial condition stated above. To this end, we use the Liouville transformation

$$\xi = \int_0^r \sqrt{n(r)}\, dr, \qquad z(\xi) = [n(r)]^{1/4} v_l(r) \tag{3.7}$$

to transform (3.5) to

$$\ddot{z} + (k^2 - p(\xi))z = 0 \tag{3.8}$$

where

$$p(\xi) = \frac{1}{4}\frac{\ddot{n}(r)}{[n(r)]^2} - \frac{5}{16}\frac{[\dot{n}(r)]^2}{[n(r)]^3} + \frac{l(l+1)}{r^2 n(r)}. \tag{3.9}$$

Note that since $n(r) > 0$ for $r \geqq 0$ and $n(r)$ is in Hölder class $C^{2,\alpha}$, the transformation (3.7) is invertible and $p(\xi)$ is well defined and Hölder continuous for $r > 0$. In order to apply known asymptotic estimates given in [13], we rewrite (3.8) in the form

$$\ddot{z} + \left(k^2 - \frac{l(l+1)}{n(0)r^2} - g(r)\right)z = 0 \tag{3.10}$$

where

$$g(r) = \frac{l(l+1)}{r^2 n(r)} - \frac{l(l+1)}{r^2 n(0)} + \frac{1}{4}\frac{\ddot{n}(r)}{[n(r)]^2} - \frac{5}{16}\frac{[\dot{n}(r)]^2}{[n(r)]^3}. \tag{3.11}$$

From (3.11) and the fact that $n(r) = 1$ for $r \geqq a$, we have that

$$\int_1^\infty |g(r)|\, dr < \infty, \qquad \int_0^1 r|g(r)|\, dr < \infty. \tag{3.12}$$

Having rewritten (3.5) in the form (3.10), (3.11), we can now use Theorem 6.1 [13, p. 450] and some straightforward estimates (cf. [13, p. 210]) to deduce the asymptotic estimate

$$\begin{aligned} z(\xi) &= \sqrt{\frac{\pi\xi}{2k}}\, J_\nu(k\xi) + O\left(\frac{\ln k}{k^2}\right) \\ &= \frac{1}{k}\cos\left(k\xi - \frac{1}{2}\nu\pi - \frac{1}{4}\pi\right) + O\left(\frac{\ln k}{k^2}\right) \end{aligned} \tag{3.13}$$

where $0 < \xi < \infty$, J_ν denotes a Bessel function of order ν, and

$$\nu = \sqrt{\frac{l(l+1)}{n(0)} + \frac{1}{4}}. \tag{3.14}$$

Furthermore, the estimate (3.13) can be differentiated with respect to ξ, the error estimate remaining the same. From (3.7), (3.13), we finally conclude that

$$v_l(r) = \frac{1}{k[n(r)]^{1/4}}\cos\left(k\int_0^r \sqrt{n(r)}\, dr - \frac{1}{2}\nu\pi - \frac{1}{4}\pi\right) + O\left(\frac{\ln k}{k^2}\right) \tag{3.15}$$

where (3.15) can be differentiated with respect to r with the same error estimate.

We now return to the interior acoustic transmission problem (3.2), (3.3) and note that (3.6) will be a nontrivial solution provided there exists a nontrivial solution a_l, b_l of the homogeneous algebraic system

$$
\begin{aligned}
&a_l\left(\frac{1}{a}\, v_l(a)\right) - b_l j_l(ka) = 0, \\
&a_l \frac{d}{dr}\left(\frac{1}{r}\, v_l(r)\right)\Big|_{r=a} - b_l \frac{d}{dr}\,(j_l(kr))|_{r=a} = 0.
\end{aligned} \tag{3.16}
$$

But (3.16) has a nontrivial solution provided that the determinant of the coefficients vanishes. Noting that

$$
j_l(kr) = \frac{1}{kr}\cos\left(kr - \frac{1}{2}\, l\pi - \frac{1}{2}\,\pi\right) + O\left(\frac{1}{k^2}\right) \tag{3.17}
$$

where (3.17) can be differentiated with respect to r, we see from (3.15), (3.19), and the addition formula for the sine function that this determinant vanishes for k large provided

$$
\sin\left[k\left(\int_0^a \sqrt{n(r)}\, dr - a\right) - \frac{1}{2}\left(\nu - l - \frac{1}{2}\right)\right] + O\left(\frac{\ln k}{k}\right) = 0. \tag{3.18}
$$

A sufficient condition for (3.18) to be valid for a discrete set of values of k is that $n(r) > 1$ for $0 \leqq r < a$, i.e., $\varepsilon(r) > \varepsilon_0$ for $0 \leqq r < a$. We could equally well have chosen $\varepsilon(r)$ such that $\varepsilon(r) < \varepsilon_0$ for $0 \leqq r < a$. In either case, we have that there is an infinite set of values of $k = \omega\sqrt{\varepsilon_0\mu_0}$ such that there exists a nontrivial solution a_l, b_l of (3.16). Summarizing our results, we have the following theorem.

THEOREM 3.1. *Assume that $\sigma(x)$ is identically zero, $\varepsilon(x) = \varepsilon(r)$ is spherically stratified, and $\varepsilon(r) > \varepsilon_0$ for $0 \leqq r < a$ or $\varepsilon(r) < \varepsilon_0$ for $0 \leqq r < a$. Then there exists an infinite set of frequencies ω such that $\mathcal{F}$ is not complete in Y.*

4. A complete set of far-field patterns. In the previous section of this paper, we showed that in general the set $\mathcal{F}$ is not complete in Y if $\sigma = 0$. In this final section, we will show that, if from each $F = F(\hat{x}; \hat{\alpha}, p) \in \mathcal{F}$ we subtract the electric far-field pattern $F_\lambda = F_\lambda(\hat{x}; \hat{\alpha}, p)$ corresponding to the solution of

$$
\operatorname{curl} E_\lambda - i\omega\mu_0 H_\lambda = 0, \qquad \operatorname{curl} H_\lambda + i\omega\varepsilon_0 E_\lambda = 0 \tag{4.1}
$$

in the exterior of $B = \{x : |x| < a\}$ such that $E_\lambda(x) = E^i(x) + E^s_\lambda(x)$, $H_\lambda(x) = H^i(x) + H^s_\lambda(x)$ with $\{E^i, H^i\}$ given by (2.2), $\{E^s_\lambda, H^s_\lambda\}$ satisfies the Silver–Müller radiation condition, and on ∂B

$$
[\hat{x}, \operatorname{curl} H_\lambda] = \lambda[\hat{x}, [\hat{x}, H_\lambda]] \tag{4.2}
$$

where $\operatorname{Im}\lambda > 0$, then the resulting class of far-field patterns is complete in Y for $\sigma(x) \geqq 0$. That is, the eigenvalues of the interior transmission problem no longer destroy completeness when $\sigma = 0$. We note that for the direct scattering problem for acoustic waves, the idea of eliminating eigenvalues by introducing an artificial boundary on which an auxiliary function satisfies an impedance boundary condition was first introduced by Ursell in 1973 [15].

We draw the reader's attention to the fact that the sign of $\operatorname{Im}\lambda$ given above is the reverse of that usually given to establish existence and uniqueness for the exterior impedance boundary value problem (4.1), (4.2) [4]. However, (4.1), (4.2) can be reformulated as an integral equation of the form $(\mathbf{I} + \mathbf{T}(\lambda))f = g$, where, except for

possibly a discrete set of values of λ, $\mathbf{T}(\lambda)$ is compact, $\mathbf{T}(\lambda)$ is a meromorphic operator valued function of λ and $\mathbf{I}+\mathbf{T}(\lambda)$ is invertible for $\operatorname{Im}\lambda<0$ [4]. Hence, by the analytic Fredholm theorem [9] and the results of [4], there exists a solution to (4.1), (4.2) for $\operatorname{Im}\lambda>0$, with the possible exception of a countable set of values of λ. We note that an explicit solution can also be found by the method of separation of variables. We choose λ such that λ is not in this exceptional set (which is possible in theory since the exceptional set is countable) and define the electric far-field pattern $F_\lambda(\hat{x};\hat{\alpha},p)$ from the electric field determined by the unique solution of the above-mentioned integral equation.

We now define the set $\mathscr{F}_\lambda$ of electric far-field patterns by

$$\mathscr{F}_\lambda=\{F(\hat{x};\hat{\alpha}_n,\hat{e}_l)-F_\lambda(\hat{x};\hat{\alpha}_n,\hat{e}_l)\colon l=1,2,3,\ n=1,2,3,\cdots\}$$

where $\hat{\alpha}_n$ and $\hat{e}_l$ are as defined in § 2 of this paper. We then have the following theorem.

THEOREM 4.1. *For any $\sigma(x)\geqq 0$ and $\varepsilon(x)\geqq 0$, the set $\mathscr{F}_\lambda$ is complete in Y.*

Proof. It suffices to show that if

$$\int_{\partial\Omega}(F(\hat{x};\hat{\alpha}_n,\hat{e}_l)-F_\lambda(\hat{x};\hat{\alpha}_n,\hat{e}_l))\cdot\overline{g(\hat{x})}\,ds\,(\hat{x})=0 \tag{4.3}$$

for some $g\in Y$ and $l=1,2,3$, $n=1,2,\cdots$, then $g(\hat{x})$ is identically zero for $\hat{x}\in\partial\Omega$. Since the set $\{\hat{\alpha}_n\}$ is dense on $\partial\Omega$, (4.3) implies that

$$\int_{\partial\Omega}(F(\hat{x};\hat{\alpha},p)-F_\lambda(\hat{x};\hat{\alpha},p))\cdot\overline{g(\hat{x})}\,ds\,(\hat{x})=0 \tag{4.4}$$

for all $\hat{\alpha}\in\partial\Omega$ and $p\in R^3$. From reciprocity (Theorem 2.1 of this paper and Theorem 3.1 of [1]) we can now conclude as in Theorem 2.2 that

$$\int_{\partial\Omega}(F(\hat{x};\hat{\alpha},h(\hat{\alpha}))-F_\lambda(\hat{x};\hat{\alpha},h(\hat{\alpha}))\,ds\,(\hat{\alpha})=0 \tag{4.5}$$

for all $\hat{x}\in\partial\Omega$ where $h(\hat{\alpha})=\overline{g(-\hat{\alpha})}$.

Now define the electromagnetic Herglotz pair $\{E_0,H_0\}$ by (2.22). Let $\{E_0^s,H_0^s\}$ be the scattered field such that $\{E_0+E_0^s,H_0+H_0^s\}$ is a solution of (2.1) and $\{E_{0\lambda}^s,H_{0\lambda}^s\}$ be the scattered field such that $\{E_0+E_{0\lambda}^s,H_0+H_{0\lambda}^s\}$ is a solution of (4.1) where $H_{0\lambda}=H_0+H_{0\lambda}^s$ satisfies (4.2). Then from (4.5) we see that E_0^s and $E_{0\lambda}^s$ have the same electric far-field pattern. Defininining $E_1=E_0+E_0^s$, $H_1=H_0+H_0^s$, and $E_{0\lambda}=E_0+E_{0\lambda}^s$, we see (Corollary 4.10 of [3]) that $E_1(x)=E_{0\lambda}(x)$ for $x\in R^3\backslash B$, and hence $H_1(x)=H_{0\lambda}(x)$ for $x\in R^3\backslash B$. This implies that $\{E_1,H_1\}$ is a solution of the following homogeneous interior impedance boundary value problem: Determine $E_1\in C^1(B)\cap C(\bar{B})$, $H_1\in C^1(B)\cap C(\bar{B})$ such that

$$\begin{aligned}&\operatorname{curl}E_1-i\omega\mu_0H_1=0\\&\operatorname{curl}H_1+i\omega\varepsilon(x)E_1=\sigma(x)E\end{aligned}\qquad\text{in } B, \tag{4.6}$$

$$[\hat{x},\operatorname{curl}H_1]=\lambda[\hat{x},[\hat{x},H_1]]\quad\text{on }\partial B. \tag{4.7}$$

We note again that if $\sigma(x)$ and $\varepsilon(x)$ are in Hölder class $C^{2,\alpha}$, then $E_1\in C^2(B)$ and $H_1\in C^2(B)$ [5]. Our aim is to now use (4.6) and (4.7) to show that the tangential components of E_1 and H_1 are identically zero and to then conclude that the Herglotz kernel g is identically zero.

From (4.7) and the identity $[\hat{x},[\hat{x},H_1]]=(\hat{x}\cdot H_1)\hat{x}-H_1$ we have that on ∂B

$$(\hat{x}\cdot H_1)\hat{x}-H_1=\frac{1}{\lambda}[\hat{x},\operatorname{curl}H_1] \tag{4.8}$$

and hence

$$\operatorname{Im}\int_{\partial B}(\hat{x}, H_1, \operatorname{curl}\bar{H}_1)\,ds = -\operatorname{Im}\frac{1}{\lambda}\int_{\partial b}(\hat{x}, [\hat{x}, \operatorname{curl} H_1], \operatorname{curl}\bar{H}_1)\,ds$$

$$= \operatorname{Im}\frac{1}{\lambda}\int_{\partial B}|[\hat{x}, \operatorname{curl} H_1]|^2\,ds \tag{4.9}$$

$$\leqq 0.$$

On the other hand, since $\operatorname{div} E_1 = 0$ on ∂B, we have from the first vector Green's theorem that

$$\operatorname{Im}\int_{\partial B}(\hat{x}, H_1, \operatorname{curl}\bar{H}_1)\,ds = \operatorname{Im}\frac{\varepsilon_0}{\mu_0}\int_{\partial B}(\hat{x}, \operatorname{curl} E_1, \bar{E}_1)\,ds$$

$$= -\operatorname{Im}\frac{\varepsilon_0}{\mu_0}\int_B\int(\bar{E}_1\cdot\Delta E_1 + |\operatorname{curl} E_1|^2 + |\operatorname{div} E_1|^2)\,dx \tag{4.10}$$

$$= -\frac{\varepsilon_0}{\mu_0}\operatorname{Im}\int_B\int\bar{E}_1\cdot\Delta E_1\,dx.$$

Since $E_1 \in C^2(B)$, we have from (4.6) that

$$\Delta E_1 - \operatorname{grad}\operatorname{div} E_1 + \omega^2\mu_0\varepsilon(x)E_1 + i\omega\mu_0\sigma(x)E_1 = 0 \tag{4.11}$$

and hence

$$\operatorname{Im}\int_B\int\bar{E}_1\cdot\Delta E_1\,dx = \operatorname{Im}\left(\int_B\int(\bar{E}_1\cdot\operatorname{grad}\operatorname{div} E_1 - \omega^2\mu_0\varepsilon(x)|E_1|^2)\,dx\right)$$

$$-\omega\mu_0\int_B\int\sigma(x)|E_1|^2\,dx$$

$$= \operatorname{Im}\int_B\int|\operatorname{div} E_1|^2\,dx - \omega\mu_0\int_B\int\sigma(x)|E_1|^2\,dx \tag{4.12}$$

$$= -\omega\mu_0\int_B\int\sigma(x)|E_1|^2\,dx$$

$$\leqq 0.$$

From (4.10) and (4.12) we have that

$$\operatorname{Im}\int_{\partial B}(\hat{x}, H_1, \operatorname{curl}\bar{H}_1)\,ds \geqq 0 \tag{4.13}$$

and hence from (4.9) we see that $[\hat{x}, \operatorname{curl} H_1] = 0$ on $\partial\Omega$, i.e., $[\hat{x}, E_1] = 0$ on $\partial\Omega$. From the impedance boundary condition (4.7) we now also have that $[\hat{x}, H_1] = 0$ on $\partial\Omega$, i.e., $[\hat{x}, \operatorname{curl} E_1] = 0$ on $\partial\Omega$.

We now show that g is identically zero, thus concluding the proof of the theorem. From Theorem 4.1 of [3], we have that for $x \in R^3\setminus\bar{B}$,

(4.14)

$$\operatorname{curl}\int_{\partial B}[\hat{y}, E_0(y)]\Phi(x, y)\,ds(y) + \frac{1}{k^2}\operatorname{curl}\operatorname{curl}\int_{\partial B}[\hat{y}, \operatorname{curl} E_0(y)]\Phi(x, y)\,ds(y) = 0$$

where $\Phi(x, y)$ is the radiating fundamental solution of the Helmholtz equation. But, from the previous paragraph, we have that $[\hat{y}, E_0(y)] = -[\hat{y}, E_0^s(y)]$ and $[\hat{y}, \operatorname{curl} E_0(y)] = -[\hat{y}, \operatorname{curl} E_0^s(y)]$ for $y \in \partial B$. Hence, from Theorem 4.5 of [3], we have that the left-hand side of (4.14) is equal to $-E_0^s(x)$ for $x \in R^3 \setminus \bar{B}$, i.e., $E_0^s(x) = 0$ for $x \in R^3 \setminus \bar{B}$. Thus, $[\hat{y}, E_0(y)] = 0$ and $[\hat{y}, \operatorname{curl} E_0(y)] = 0$ for $y \in \partial B$, and from Theorem 4.1 of [3] we now have that $E_0(x) = 0$ for $x \in B$ and, by unique continuation, for $x \in R^3$. We can now conclude that g is identically zero and the proof of the theorem is complete.

REFERENCES

[1] T. S. ANGELL, D. COLTON, AND R. KRESS, *Far field patterns and inverse scattering problems for imperfectly conducting obstacles*, Math. Proc. Cambridge Philos. Soc., 20 (1989), pp. 1472–1483.

[2] D. COLTON, A. KIRSCH, AND L. PÄIVÄRINTA, *Far field patterns for acoustic waves in an inhomogeneous medium*, SIAM J. Math. Anal., 20 (1989), pp. 1472–1483.

[3] D. COLTON AND R. KRESS, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.

[4] ———, *The impedance boundary value problem for the time harmonic Maxwell equations*, Math. Methods Appl. Sci., 3 (1981), pp. 475–487.

[5] ———, *Time harmonic electromagnetic waves in an inhomogeneous medium*, Proc. Roy. Soc. Edinburgh, to appear.

[6] D. COLTON AND P. MONK, *The inverse scattering problem for time harmonic acoustic waves in an inhomogeneous medium*, Quart. J. Mech. Appl. Math., 41 (1988), pp. 97–125.

[7] ———, *The inverse scattering problem for time harmonic acoustic waves in an inhomogeneous medium: numerical experiments*, IMA J. Appl. Math., 42 (1989), pp. 77–95.

[8] D. COLTON AND L. PÄIVÄRINTA, *Far field patterns and the inverse scattering problem for electromagnetic waves in an inhomogeneous medium*, Math. Proc. Cambridge Philos. Soc., 103 (1988), pp. 561–575.

[9] N. DUNFORD AND J. SCHWARTZ, *Linear Operators, Vol.* I, John Wiley, New York, 1958.

[10] B. FRIEDMAN, *Propagation in a non-homogeneous atmosphere*, in Theory of Electromagnetic Waves, M. Kline, ed., Dover, New York, 1965, pp. 317–350.

[11] D. S. JONES, *Acoustic and Electromagnetic Waves*, Clarendon Press, Oxford, 1986.

[12] C. MÜLLER, *Foundations of the Mathematical Theory of Electromagnetic Waves*, Springer-Verlag, New York, 1969.

[13] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[14] W. TABBARA, B. DUCHÊNE, C. PICHOT, D. LESSELIER, L. CHOMMELOUX, AND N. JOACHIMOWICZ, *Diffraction tomography: contribution to the analysis of some applications in microwaves and ultrasonics*, Inverse Problems, 4 (1988), pp. 305–331.

[15] F. URSELL, *On the exterior problems of acoustics*, Proc. Cambridge Philos. Soc., 74 (1973), pp. 117–125.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1550–1562, November 1990
©1990 Society for Industrial and Applied Mathematics
010

ON THE INTEGER TRANSLATES OF A COMPACTLY SUPPORTED FUNCTION: DUAL BASES AND LINEAR PROJECTORS*

ASHER BEN-ARTZI† AND AMOS RON‡

Abstract. Given a multivariate compactly supported function ϕ, linear projectors to the space $S(\phi)$ spanned by its integer translates are discussed here. These projectors are constructed with the aid of a dual basis for the integer translates of ϕ, hence under the assumption that these translates are linearly independent. The main result shows that the linear functionals of the dual basis are local, hence making it possible to contruct local linear projectors onto $S(\phi)$. A scheme for the construction of such local projectors is then discussed for a general compactly supported function.

In the second part of the paper these observations are applied to piecewise-polynomials and piecewise-exponentials to obtain a necessary and sufficient condition for a quasi interpolant to be a projector. The results of that part extend and refine recent constructions of dual bases and linear projectors for polynomial and exponential box splines.

Key words. exponentials, polynomials, multivariate, multivariate splines, uniform mesh, regular grids, integer translates, dual bases, linear projectors

AMS(MOS) subject classifications. primary 41A15, 41A63; secondary 41A05, 46A22, 65D07

1. Introduction.

Let ϕ be a complex-valued compactly supported continuous function defined on $\mathbb{R}^s$, and E^α the *shift* operator

$$E^\alpha f = f(\cdot + \alpha).$$

Associated with ϕ is the semidiscrete convolution operator

$$\phi* : c \mapsto \phi * c := \sum_{\alpha \in \mathbb{Z}^s} c(\alpha) E^{-\alpha} \phi, \tag{1.1}$$

which maps the set

$$\mathcal{C} := \{c : \mathbb{Z}^s \to \mathbb{C}\}$$

of all complex-valued sequences defined on the multi-integers to the set

$$S(\phi) := \operatorname{ran} \phi*$$

of the functions spanned by the integer translates of ϕ.

*Received by the editors May 8, 1989; accepted for publication (in revised form) December 1, 1989.

†Laboratory for Mathematics and Statistics, University of California at San Diego, La Jolla, California 92093. The work of this author was partially supported by a Dr. Chaim Weizmann fellowship for scientific research.

‡Center for the Mathematical Sciences and Computer Sciences Department, 1210 West Dayton Street, University of Wisconsin-Madison, Madison, Wisconsin 53706.

In multivariate spline approximation, $S(\phi)$ is of potential use as a space of approximants for a larger function space (e.g., $C(\mathbb{R}^s)$). The operator $\phi*$ is exploited in the derivation of explicit approximation schemes from $S(\phi)$. Such a scheme may result in a linear projector onto $S(\phi)$ (cf. [B1] and [BF] for the construction of linear projectors for *univariate* and *tensor product* splines). Since the construction of linear projectors in the multivariate case is usually quite involved, we are satisfied with the so-called quasi-interpolation schemes, which yield the same approximation order.

In case the integer translates of ϕ are globally linearly independent (i.e., $\ker(\phi*) = 0$), a natural way to define a linear projector onto $S(\phi)$ is with the aid of a linear functional λ that satisfies

$$\lambda(E^\alpha \phi) = \delta_{\alpha,0}, \qquad \alpha \in \mathbb{Z}^s.$$

For then the functionals $\Lambda := \{\lambda E^\alpha\}_\alpha$ form a dual basis for $\Phi := \{E^{-\alpha}\phi\}_\alpha$, and a linear projector Ψ can than be defined in the usual way:

$$\Psi = \sum_{\alpha \in \mathbb{Z}^s} E^{-\alpha}\phi \; \lambda E^\alpha.$$

For the analysis of the approximation properties of the projector, its *localness* is important: a projector Ψ is termed "local" if for every compact $A \subset \mathbb{R}^s$ there exists a compact $B \subset \mathbb{R}^s$ such that $\Psi(f)_{|A}$ is determined by $f_{|B}$. The existence of a local projector is guaranteed in case the generator λ of the dual basis Λ is local. The construction of local linear projectors is facilitated if it is assumed that the integer translates of ϕ are *locally linearly independent*, i.e., the condition

$$(\phi * c)_{|A} = 0 \text{ and } \operatorname{supp} E^{-\alpha}\phi \cap A \neq \emptyset \Longrightarrow c(\alpha) = 0, \tag{1.2}$$

for every open $A \subset \mathbb{R}^s$. It is one of the main themes of this paper to show that local projectors can be constructed even under the weaker assumption of global linear independence. As a matter of fact, that observation neither makes use of the shift-invariance of the space $S(\phi)$ nor of the fact that ϕ is a function.

THEOREM 1.3. *Let $\Phi = \{\phi_\alpha\}_{\alpha \in \mathbb{Z}^s}$ be a locally finite collection of (globally) linearly independent compactly supported distributions in $\mathcal{D}'(\mathbb{R}^s)$. Then each functional λ_α in the (algebraic) dual basis $\Lambda = \{\lambda_\alpha\}_{\alpha \in \mathbb{Z}^s}$ of Φ is local. Precisely, for every $\alpha \in \mathbb{Z}^s$ there exists a compact $B_\alpha \subset \mathbb{R}^s$ such that $\lambda_\alpha(f) = 0$ whenever* $\operatorname{supp} f \cap B_\alpha = \emptyset$.

Here, *locally finite* means that only finitely many elements have some of their support in any given compact set.

In §2 we prove Theorem 1.3 and employ this result to show that the assumption of linear independence of the integer translates of ϕ is already sufficient to allow the existence of a dual basis based on a linear functional λ of point-evaluation at a finite set of $\mathbb{R}^s$. The proof of the theorem also gives information about the diameter of $\operatorname{supp} \lambda$ which in the case of local linear independence coincides with the standard result.

The construction of local projectors is then discussed in §3. There we take ϕ to be an arbitrary function whose translates are linearly independent, and, based on a new extended notion of a quasi interpolant, provide a necessary and sufficient condition for a quasi interpolant to be a linear projector. With the aid of this observation, we then describe a general scheme for the derivation of local projectors onto $S(\phi)$.

In the next two sections we examine the piecewise-exponential case (which contains the piecewise-polynomial case). Section 4 is devoted to a brief discussion of some known methods for constructing quasi interpolants for piecewise-exponentials. These results, together with observations from §3, are used in the §5 where we show that for piecewise-exponentials a slightly stronger sense of local linear independence is sufficient to imply that every quasi interpolant is also a linear projector. We conclude that section with a review of the constructions of linear projectors in [DM1], [DM2], [J1], and [J2] providing thereby new proofs and extensions to these results.

Finally, we provide in the last section an example which demonstrates the fact that a certain stronger (still natural) version of Theorem 1.3 fails to hold.

2. Linear projectors are local. In this section we prove Theorem 1.3 and discuss some of its applications.

The following is an equivalent form of Theorem 1.3 which is slightly more convenient for the proof employed.

THEOREM 2.1. *Let $\Phi = \{\phi_\alpha\}_{\alpha\in\mathbb{Z}^s}$ be a locally finite collection of globally linearly independent compactly supported distributions in $\mathcal{D}'(\mathbb{R}^s)$. Then there exists a ball*

$$B := \{x : \|x\| \le L\} \tag{2.2}$$

such that if $f = \sum_{\alpha\in\mathbb{Z}^s} c(\alpha)\phi_\alpha$ satisfies $\operatorname{supp} f \cap B = \emptyset$ then $c(0) = 0$.

Indeed, Theorem 1.3 readily follows from Theorem 2.1: assume (without loss) that $\alpha = 0$ in Theorem 1.3. For every $f = \sum_{\alpha\in\mathbb{Z}^s} c(\alpha)\phi_\alpha \in \operatorname{span}\Phi$, $\lambda_0(f) = c(0)$. So, if we assume Theorem 2.1 and choose B_0 of Theorem 1.3 to be B in (2.2), then whenever $\operatorname{supp} f \cap B = \emptyset$, Theorem 1.3 implies $\lambda_0(f) = c(0) = 0$.

We postpone the proof of Theorem 2.1 to the end of this section, but discuss some of its applications now.

Suppose that $\Phi \subset C(\mathbb{R}^s)$ (or any other space on which point-evaluation is well defined) and choose $\alpha \in \mathbb{Z}^s$. Let B_α be the ball associated with α (by Theorem 1.3). Define

$$\nu_\alpha := \{\beta \in \mathbb{Z}^s : \operatorname{supp}\phi_\beta \cap B_\alpha \neq \emptyset\}. \tag{2.3}$$

Since Φ is locally finite, the set ν_α is finite. Defining

$$S(\Phi) := \operatorname{span}\Phi, \qquad S_\alpha(\Phi) := \operatorname{span}\{\phi_{\beta|B_\alpha} : \beta \in \nu_\alpha\},$$

Theorem 1.3 implies that any extension $\mu_\alpha \in C(\mathbb{R}^s)^*$ of the restricted linear functional $\lambda_{\alpha|S_\alpha(\Phi)}$ is also an extension of λ_α, provided that $\operatorname{supp}\mu_\alpha \subset B_\alpha$. Now, $S_\alpha(\Phi)$ is a finite-dimensional subspace of $C(B_\alpha)$, B_α being bounded, and hence there are various ways available for representing the restriction of λ_α to $S_\alpha(\Phi)$; e.g., we may choose a set $b_\alpha \subset B_\alpha$ of cardinality $\#\nu_\alpha$, which is *total* for $S_\alpha(\Phi)$ (i.e., no element in $S_\alpha(\Phi)\backslash 0$ vanishes on b_α). Then there is a unique linear combination $\mu_\alpha = \sum_{x\in b_\alpha} c(x)[x]$ satisfying

$$\mu_\alpha(f) = \lambda_\alpha(f) \quad \forall f \in S(\Phi), \tag{2.4}$$

where $[x]$ is the functional of point-evaluation at x. Thus we conclude the following result.

COROLLARY 2.5. *Assume that* $\Phi \subset C(\mathbb{R}^s)$. *Then there exists a projector*

$$\Psi : C(\mathbb{R}^s) \to S(\Phi) : f \mapsto \sum_{\alpha \in \mathbb{Z}^s} \mu_\alpha(f)\phi_\alpha \tag{2.6}$$

such that each μ_α *is supported on a finite set.*

In the special case of interest, viz., when $\phi_\alpha = E^{-\alpha}\phi$, we have

$$\sum_{x \in b} c(x)\phi(x+\alpha) = \mu(E^\alpha \phi) = \begin{cases} 1, & \alpha = 0, \\ 0, & \alpha \neq 0 \end{cases}$$

(with $\phi := \phi_0$, $\mu := \mu_0 = \sum_{x \in b} c(x)[x]$, and $b := b_0$). This proves the following.

COROLLARY 2.7. *Let* ϕ *be a compactly supported continuous function whose integer translates are globally linearly independent. Then there exists a finite linear combination* ψ *of (real) translates of* ϕ *satisfying*

$$\psi_{|\mathbb{Z}^s} = \delta_{\alpha,0}.$$

Proof of Theorem 2.1. Assume, to the contrary, that such a ball B does not exist. For every positive integer n, let B_n be the open ball centered at the origin with radius n and define

$$\nu_n := \{\alpha \in \mathbb{Z}^s : \operatorname{supp} \phi_\alpha \cap B_n \neq \emptyset\}. \tag{2.8}$$

Let M_n be the linear space of all sequences c defined on ν_n and satisfying

$$B_n \cap \operatorname{supp}\left(\sum_{\alpha \in \nu_n} c(\alpha)\phi_\alpha\right) = \emptyset. \tag{2.9}$$

Since Φ is locally finite, every ν_n is finite, and hence every M_n is finite-dimensional. On the other hand, no M_n is trivial, since by our assumption each M_n must contain at least one element satisfying $c(0) \neq 0$. To obtain the desired contradiction, we will show that there is a nontrivial sequence in $\mathcal{C}$ whose restriction to each ν_n lies in M_n. Since the union of the sets $\{\nu_n\}_n$ is $\mathbb{Z}^s$, such a sequence c induces a nontrivial vanishing combination of Φ thus contradicting the linear independence of the elements in Φ.

For this purpose, define $M_0 := \mathbb{C}$ and for all nonnegative integers $m \geq n > 0$, let T_n^m be the restriction map from M_m to M_n (with $T_0^n : c \mapsto c(0)$ and T_0^0 the identity mapping). Clearly,

$$T_0^m \neq 0 \quad \forall\, m. \tag{2.10}$$

Defining $K_n \subset M_n$ by

$$K_n := \bigcap_{m \geq n} \operatorname{ran} T_n^m,$$

we note that the condition

$$K_n \neq 0$$

is necessary (but apparently not sufficient) for the existence of a sequence c whose restriction to each ν_n lies in M_n. Indeed, we claim that, for each $n \geq 0$, $K_n \neq 0$.

For this, note that, for every $n \le k \le m$,

$$T_n^m = T_n^k T_k^m, \tag{2.11}$$

while $\{\operatorname{ran} T_n^m\}_{m \ge n}$ is a decreasing sequence of finite-dimensional vector spaces. Therefore, for all sufficiently big k and m, $T_n^m = T_n^k$, and hence $K_n = \operatorname{ran} T_n^m$ for all sufficiently big m. Finally, since $T_0^m = T_0^n T_n^m$, then (2.10) implies that $T_n^m \neq 0$; hence so is K_n. The fact that, for sufficiently large m, $K_n = \operatorname{ran} T_n^m$ implies that we can find m such that

$$K_j = \operatorname{ran} T_j^m, \qquad j = n, n+1,$$

and thus

$$K_n = T_n^{n+1} T_{n+1}^m M_m = T_n^{n+1} K_{n+1}. \tag{2.12}$$

We can now complete the proof as follows: let n be arbitrary, and let $c_n \in K_n$ be an arbitrary nontrivial element. Invoking (2.12), we may choose $c_{n+1} \in K_{n+1}$ whose restriction to ν_n is c_n. Again, (2.12) can be employed to provide $c_{n+2} \in K_{n+2}$ whose restriction to ν_{n+1} is c_{n+1}. Proceeding in this manner, we obtain $(c_m \in K_m)_{m=n}^{\infty}$ such that $T_n^m c_m = c_n$. This gives rise to a sequence $c \in \mathcal{C}$ satisfying

$$B_m \cap \operatorname{supp}\left(\sum_{\alpha \in \nu_m} c(\alpha)\phi_\alpha\right) = \emptyset \quad \forall m,$$

while c is nontrivial since its restriction c_n to ν_n is nontrivial.

We therefore conclude that the distributions in Φ are globally linearly dependent, in contradiction to the assumptions of the theorem. □

3. Linear projectors and quasi interpolation. Throughout this section we assume that ϕ is a compactly supported continuous function whose integer translates are globally linearly independent, and $A \subset \mathbb{R}^s$ is an open bounded set which satisfies for every $c \in \mathcal{C}$ the condition

$$(\phi * c)_{|A} = 0 \implies c(0) = 0. \tag{3.1}$$

The existence of such an A was proved in Theorem 1.3.

Given a linear functional $\lambda : C(\mathbb{R}^s) \to \mathbb{C}$, we examine here conditions that guarantee that the operator

$$Q_\lambda := \phi * \Lambda(\cdot)$$

is a projector. Here

$$\Lambda : C(\mathbb{R}^s) \to \mathcal{C} : f \mapsto (\lambda E^\alpha f)_{\alpha \in \mathbb{Z}^s}.$$

Our aim is to generate a space $F \subset S(\phi)$ (which replaces the missing polynomial space usually associated with a piecewise-polynomial ϕ) that is of help in the identification of a projector Q_λ. We use the notation $f_| := f_{|\mathbb{Z}^s}$, and let $\phi *'$ stand for the semidiscrete convolution operator from $C(\mathbb{R}^s)$ to $S(\phi)$ defined by

$$\phi *' f := \phi * f_| = \sum_{\alpha \in \mathbb{Z}^s} f(\alpha) E^{-\alpha}\phi. \tag{3.2}$$

Finally, $\phi \,'*'\, f$ denotes the discrete convolution $\phi_| * f_|$.

THEOREM 3.3. *Let F be a subspace of $S(\phi)$ satisfying*

$$F_{|A} = S(\phi)_{|A}, \tag{3.4}$$

and assume that $\operatorname{supp}\lambda \subset A$. *Then the following conditions are equivalent:*

(a) $$Q_\lambda(f) = f \quad \forall f \in F.$$

(b) Q_λ *is a projector, i.e.,*

$$Q_\lambda(f) = f \quad \forall f \in S(\phi).$$

Proof. The implication (b) $\Longrightarrow$ (a) is trivial. For the converse, it is necessary and sufficient to prove that

$$\lambda(E^{-\alpha}\phi) = \begin{cases} 1, & \alpha = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Fix α and let $f \in F$ be such that $f_{|A} = (E^{-\alpha}\phi)_{|A}$; then, since $Q_\lambda(f) = f$,

$$Q_\lambda(f)_{|A} = E^{-\alpha}\phi_{|A}.$$

Now, $E^{-\alpha}\phi = \phi * \delta_\alpha$ (with $\delta_\alpha(\beta) := \delta_{\alpha,\beta}$), while $Q_\lambda(f) = \phi * \Lambda(f)$, and therefore, by (3.1), we must have $\lambda(f) = \Lambda(f)(0) = \delta_\alpha(0)$. On the other hand, λ is supported on A and therefore since $E^{-\alpha}\phi$ and f coincide on A we conclude $\lambda(E^{-\alpha}\phi) = \lambda(f) = \delta_\alpha(0)$. □

The assumption $\operatorname{supp}\lambda \subset A$ in the theorem is essential, as shown by the following simple example.

Example 3.5. Let ϕ be the univariate hat function supported on $[0, 2]$. Let $F = \pi_1$, $\lambda = \frac{1}{2}([\frac{1}{2}] + [1\frac{1}{2}])$. Then Q_λ reproduces π_1, and $F = \pi_1$ also satisfies (3.4) with A being any subset of $[0, 1]$ or $[1, 2]$. Yet, λ is not supported in any of these A's and therefore Q_λ is not guaranteed to be a projector. Indeed,

$$\lambda(E^{\alpha}\phi) = \begin{cases} \frac{1}{2}, & \alpha = 0, \\ \frac{1}{4}, & \alpha = \pm 1, \\ 0, & \text{otherwise.} \end{cases}$$

Needless to say, there exist projectors whose corresponding λ is supported in no admissible A; e.g., $\lambda = \frac{1}{2}(-[\frac{1}{3}] + 2[\frac{2}{3}] + 2[\frac{4}{3}] - [\frac{5}{3}])$.

We now employ the above theorem to show that, with an appropriate choice of the space F, the task of constructing linear projectors is reduced to the construction of the so-called quasi interpolants. For that, assume that F is a shift-invariant (i.e., closed under integer translates) subspace of $S(\phi)$. Then it follows [B2], [R] that F is an invariant subspace for $\phi*'$. If we further assume that F is finite-dimensional and $\phi*'$ is 1-1 on F, then $\phi*'$ induces an automorphism on F. We may then follow [BH], call this automorphism T and define a functional on F by

$$[0]T^{-1} : f \mapsto T^{-1}f(0).$$

For $\mu \in F^*$ rather than $[0]T^{-1}$, the linear independence of the integer translates of ϕ would imply that Q_μ is not the identity mapping on F and therefore no extension of such μ would yield a projector. On the other hand, for $\mu = [0]T^{-1}$ it follows [BH], that $\mu(E^\alpha f) = T^{-1}f(\alpha)$ and therefore (with Q_μ defined only on F)

$$Q_\mu(f) = f \quad \forall f \in F.$$

This leads to the following conclusion.

PROPOSITION 3.6. *For $\lambda \in V^*$, the condition*

$$\lambda(f) = [0]T^{-1}(f) \quad \forall f \in F,$$

is sufficient for Q_λ to be a quasi interpolant. Furthermore, this condition is also necessary in case $\phi*$ is injective.

If we furthermore assume that $f \mapsto f_{|A}$ is 1-1 on F, the functional μ can be extended to a functional $\lambda : C(\mathbb{R}^s)$ that is supported on A and then Theorem 3.3 would imply that Q_λ is a projector.

We summarize all these observations in the following corollary.

COROLLARY 3.7. *Let F be a finite-dimensional shift-invariant subspace of $S(\phi)$ which satisfies* (3.4). *Assume that the operator $T := \phi *'_{|F}$ is injective and consider the following conditions for a linear functional $\lambda \in C(\mathbb{R}^s)^*$:*

(a) *Q_λ is a projector.*

(b) *$\lambda(f) = [0]T^{-1}f \quad \forall f \in F$.*

Then (a)$\Longrightarrow$(b), *and the converse implication holds provided that λ is supported in A.*

The construction of a quasi interpolant using the inverse of the map T appears first in [BH] in the context of the approximation order for box splines (hence for a polynomial F), without, however, making the connection to linear projectors. The discussion above emphasizes the fact (which is by now well known) that this is (essentially) the only way to construct quasi interpolants. In particular, other constructs (cf., e.g., [SF], [DM1], [CD], [CL] for piecewise-polynomials and [DR] for piecewise-exponentials) are as a matter of fact special ways to extend $[0]T^{-1}$.

Our next task is to prove the existence of a space F which satisfies all these conditions. This purpose can be accomplished without an appeal to linear independence. First, we associate with A the set

$$\nu(A) := \nu_\phi(A) := \{\alpha \in \mathbb{Z}^s : \ A - \alpha \cap \operatorname{supp} \phi \neq \emptyset\}, \tag{3.8}$$

which consists of all integers α whose corresponding $E^{-\alpha}\phi$ has some support on A. Furthermore, we assume without loss that $\phi_| \neq 0$; otherwise ϕ can be replaced by one of its noninteger translates.

We now look for a shift-invariant space P which on the one hand interpolates correctly on $\nu(A)$ (that is, $\dim P = \#\nu(A)$ and no $p \in P\backslash 0$ vanishes identically on $\nu(A)$), while on the other hand has trivial intersection with $\ker\phi\ '*'$. Any space satisfying these two conditions will do here. In particular, we may first choose P to be a homogeneous translation-invariant polynomial space that interpolates correctly on $\nu(A)$, (cf. [BR1] for construction of such P of least degree). If $\phi\ '*'$ is not 1-1 on P, it may be replaced by a space $P_\theta := e_\theta P$, where the exponential $e_\theta : x \mapsto e^{\theta\cdot x}$ is chosen such that the *discrete* convolution $\phi\ '*' e_\theta \neq 0$. This readily implies that the operator $\phi\ '*'$ is 1-1 on $e_\theta\pi$ (and hence so is $\phi*'$). Since for every $\theta \in \mathbb{C}^s$, the space P_θ also interpolates correctly on $\nu(A)$, P_θ satisfies the required conditions.

Now define

$$F := \phi *' P_\theta.$$

Since P_θ is translation-invariant, hence shift-invariant, so is F. F is also finite-dimensional; in fact, since $\phi*'$ is 1-1 on P_θ, $\dim F = \#\nu(A)$. Moreover, by the definition of $\nu(A)$, $S(\phi)_{|A} = \operatorname{span}\{E^{-\alpha}\phi_{|A}\}_{\alpha\in\nu(A)}$, and since P_θ interpolates correctly on $\nu(A)$, we have

$$F_{|A} = S(\phi)_{|A}.$$

Finally, the discrete convolution $\phi\ '*'$ is injective on P_θ, thus induces an automorphism on that space and consequently F coincides with P_θ on $\mathbb{Z}^s$. We then conclude as follows.

THEOREM 3.9. *Let ϕ be a compactly supported continuous function, and assume that $\phi_| \neq 0$. Let A be an open subset of $\mathbb{R}^s$. Then there exists a shift-invariant subspace $F \subset S(\phi)$ satisfying*

(a) $\dim F = \#\nu(A)$;
(b) $\phi *' |_F$ *is an automorphism;*
(c) $F|_A = S(\phi)|_A$.
(d) $F_|$ *is, up to multiplication by an exponential, a homogeneous (sequence) polynomial space.*

The following scheme sketches the various steps required in the construction of linear projectors using the approach above.

SCHEME 3.10. *Let ϕ be a compactly supported continuous function whose integer translates are globally linearly independent. Check whether $\phi_| = 0$; if so replace ϕ by a translate of it. Find a subset $A \subset \mathbb{R}^s$ satisfying* (3.1)*; then*

(a) *Compute the finite set $\nu(A)$.*
(b) *Find a polynomial space P which interpolates correctly on $\nu(A)$. For that you may apply the algorithm given in* [BR1; §4.]
(c) *Find an exponential e_θ such that $\sum_{\alpha\in\mathbb{Z}^s} e_\theta(\alpha)\phi(-\alpha) \neq 0$. Define $F = \phi *' e_\theta P$. At this point you may wish to replace F by a shift-invariant subspace of it which still satisfies* (3.4).
(d) *For a given basis $f_1, \cdots, f_n$ for F, find the basis $g_1, \cdots, g_n$ for F that satisfies*

$$\phi *' g_j = f_j,\ j = 1, \cdots, n.$$

(This step can also be executed with discrete convolution, i.e., with $\phi_|$ replacing ϕ).
(e) *Define a linear functional on F by $\mu(f_j) = g_j(0)$, $j = 1, \cdots, n$, and extend μ to your favourite choice of a functional λ defined on some superspace of $S(\phi)$ and supported in A. (In case the extension is to $C(\mathbb{R}^s)$, you may choose λ to be supported on an appropriate finite subset of A with cardinality $\leq \#\nu(A)$).*
(f) *The resulting Q_λ is a linear projector.*

4. Piecewise-polynomials and piecewise-exponentials: quasi interpolation. In this section we review some methods concerning the construction of quasi interpolants for piecewise-exponentials (and in particular for piecewise-polynomials). Most of the results here are known, and the approach taken follows that of [BR2]. The discussion also makes use of various observations from [BAR], [B2], and [R].

Let H be a finite-dimensional exponential space. This means, by definition, that each function in H admits the form

$$\sum_{j=1}^{n} e_{\theta_j} p_j,\ \theta_j \in \mathbb{C}^s,\quad p_j \in \pi,\quad j = 1, \cdots, n. \tag{4.1}$$

We refer to the elements of H simply as "exponentials." Furthermore, we assume hereafter that H is translation-invariant, which implies that a basis for H is given in terms of functions of the form

$$e_\theta p, \quad p \in \pi.$$

The *spectrum* of H is the set of frequencies of all exponentials in H:

$$\operatorname{spec} H := \{\theta \in \mathbb{C}^s : \ e_\theta \in H\}. \tag{4.2}$$

Now let ϕ be a compactly supported function for which

$$H \subset S(\phi). \tag{4.3}$$

A *quasi interpolant* here means any linear map Q from some superspace V of $S(\phi)$ into $S(\phi)$ which satisfies

$$Q(f) = f \quad \forall f \in H. \tag{4.4}$$

Since H is shift-invariant, $T := \phi *' |_H$ is an endomorphism. Furthermore, under the regularity assumption

$$\widehat{\phi}(-i\theta) \neq 0 \quad \forall \theta \in \operatorname{spec} H, \tag{4.5}$$

the operator T is also an automorphism. Indeed, $\ker T$ is always shift-invariant, hence if it is not trivial, it must contain an exponential e_θ. Yet, $\phi *' e_\theta = \widehat{\phi}(-i\theta) e_\theta$ (as follows, say, from Result 4.6 below), and thus, under (4.5), $\ker T = 0$.

Proposition 3.6 shows that a careful study of the map T^{-1}, hence also of T, is essential for construction of a quasi interpolant. This study is facilitated by the following result, which is obtained by an application of Poisson's summation formula. Using the standard weak argument (i.e., applying the formula to test functions) this result is shown [BR2], to hold for an arbitrary compactly supported distribution ϕ.

RESULT 4.6. *Let $e_\theta p$ be a function in $S(\phi)$ (with $\theta \in \mathbb{C}^s$ and $p \in \pi$). Then*

$$\phi *' (e_\theta p) = \phi * (e_\theta p),$$

where the right-hand side convolution is the usual convolution between functions (or distributions).

The above result suggests that in order to find a pre-image of $e_\theta p \in H$, we may solve the convolution equation

$$\phi * (e_\theta ?) = e_\theta p.$$

Dividing both sides by e_θ and applying the Fourier transform we get

$$(E^{-i\theta}\widehat{\phi})\widehat{?} = \widehat{p}.$$

Since $\operatorname{supp} \widehat{p} = 0$ and $\widehat{\phi}(-i\theta) \neq 0$, we may divide both sides of (4.6) by $E^{-i\theta}\widehat{\phi}$ to conclude

$$\widehat{?} = \frac{\widehat{p}}{E^{-i\theta}\widehat{\phi}} = \sum_{\alpha \geq 0} \frac{D^\alpha \psi(\theta)}{\alpha!} \widehat{D^\alpha p}, \qquad \psi(x) := \frac{1}{\widehat{\phi}(-ix)};$$

hence

$$? = \sum_{\alpha \geq 0} \frac{D^\alpha \psi(\theta)}{\alpha!} D^\alpha p.$$

With $D^\alpha \psi(\theta)/\alpha!$ denoted by $a_{\theta,\alpha}$, we conclude the following.

PROPOSITION 4.7. *For $e_\theta p \in H$*

$$[0]T^{-1}(e_\theta p) = [0] \sum_{\alpha \geq 0} a_{\theta,\alpha} D^\alpha p. \tag{4.8}$$

Combining the last proposition with Proposition 3.6 we finally obtain the following.

COROLLARY 4.9. *Assume that the integer translates of ϕ are globally linearly independent. Then the condition*

$$\lambda(e_\theta p) = [0] \sum_{\alpha} a_{\theta,\alpha} D^\alpha p \quad \forall e_\theta p \in H$$

is necessary and sufficient for Q_λ to be a quasi interpolant.

5. Piecewise-polynomials and piecewise-exponentials: linear projectors. Here we combine the results of the two previous sections in the derivation of linear projectors for the piecewise-exponential space $S(\phi)$.

Retaining the notation of §4, we assume throughout this section that $H \subset S(\phi)$, and that there exists an open bounded set $A \subset \mathbb{R}^s$ for which

$$\#\nu(A) = \dim H \tag{5.1}$$

(where $\nu(A)$ is as in (3.8)). It follows that H satisfies the condition required of F in (3.4). Furthermore, the translates of ϕ are locally linearly independent on A (in the sense of (1.2)), and in particular A satisfies (3.1). With the aid of [R, Lem. 2.2] we can also easily conclude that $\phi *'$ is 1-1 on H.

Therefore, Corollary 3.7 reads as follows.

COROLLARY 5.2. *Let λ be a linear functional defined on an extension V of $S(\phi)$ and vanishing on all $f \in S(\phi)$ supported in $\mathbb{R}^s \backslash A$. If*

$$Q_\lambda(f) = f \quad \forall f \in H,$$

then Q_λ is a projector.

Equivalently, Q_λ is a projector if λ extends the linear functional $[0]T^{-1} \in H^*$. The precise values of $[0]T^{-1}$ on H were determined in Corollary 4.9, but of course many extensions (to various V's) of $[0]T^{-1}$ are available. To draw the connections between the results here and the constructions of dual bases for a box spline space in [DM1], [DM2], [J1], and [J2], we concentrate now on the case when $[0]T^{-1}$ is represented (and thus extended) with the aid of differential operators. First, we associate with every $q \in \pi$ a linear functional $q^* \in H^*$ defined by

$$q^*(f) = q(D)f(0) \quad \forall f \in H.$$

Note that for $e_\theta p \in H$

$$q^*(e_\theta p) = p^* E^\theta(q) = [0] \sum_{\alpha \geq 0} \frac{(D^\alpha q)(\theta)}{\alpha!} D^\alpha p,$$

while, on the other hand, by Corollary 4.9

$$[0]T^{-1}(e_\theta p) = [0]\sum_{\alpha\geq 0} a_{\theta,\alpha} D^\alpha p.$$

Hence we conclude the following.

COROLLARY 5.3. *Let $q \in \pi$ satisfy*

$$D^\alpha q(\theta) = \alpha!\, a_{\theta,\alpha} \quad \forall \theta \in \operatorname{spec} H, \quad |\alpha| \leq \max\{\deg p : \ e_\theta p \in H\}, \tag{5.4}$$

with $\{a_{\theta,\alpha}\}$ as in Corollary 4.9. For a space $V \supset S(\phi)$, let $\lambda \in V^*$ be an extension of $q^* \in H^*$ which vanishes on all functions in $S(\phi)$ with support in $\mathbb{R}^s \backslash A$. Then Q_λ is a projector.

In case all the integer translates of ϕ belong to H in a neighborhood of the origin, we may choose to extend q^* (at least on $S(\phi)$) to the functional $\lambda(f) = q(D)f(0)$, $f \in S(\phi)$. We note that in general condition (5.4) is sufficient but not necessary for the equality $[0]T^{-1} = q^*$.

Choosing ϕ to be a polynomial or exponential box spline, Corollary 5.3 verifies [DM1, Thm. 5.1] and [DM2, Thm. 5.1]. Note that, in contrast to [DM1] and [DM2], the approach taken here avoids the application of Poisson's summation formula at this stage, hence we need not impose any further restrictions on the polynomial q (see [J2] for a discussion of the difficulty in the application of Poisson's summation formula). Poisson's formula is still implicitly used here, since this is the key tool in the proof of Result 4.6. Nevertheless, as mentioned before, that latter result holds for any compactly supported distribution ϕ.

In case ϕ is not smooth enough at the origin, we may wish to represent $[0]T^{-1}$ by $q_\theta^* E^\theta$ with $\theta \in \mathbb{R}^s$ chosen such that $S(\phi)$ is locally in H in a neighborhood of θ. To find the connection between the various q_θ's, let $P \subset \pi$ be a space *dual* to H in the sense that the map

$$P \to H^* : q \mapsto q^*$$

is bijective (hence every $\mu \in H^*$ is uniquely represented by some q^* with $q \in P$). With $\{f_j\}_{j=1}^n$ and $\{p_j\}_{j=1}^n$ dual bases for H and P, respectively, we have

$$f = \sum_{j=1}^n p_j^*(f) f_j \quad \forall f \in H.$$

Let q be the unique polynomial in P satisfying $q^* = [0]T^{-1}$ and let $\theta \in \mathbb{R}^s$. Then

$$\begin{aligned}
[0]T^{-1}(f) = q^*(f) &= (q^* E^{-\theta})(E^\theta f) \\
&= \sum_{j=1}^n p_j^*(E^\theta f) q^*(E^{-\theta} f_j) \\
&= [\theta]\left(\sum_{j=1}^n (q(D)f_j)(-\theta) p_j(D) f\right).
\end{aligned}$$

Since H is translation-invariant and q does not vanish on spec H, $q(D)$ is injective on H; hence $\{q(D)f_j\}_{j=1}^n$ is also a basis for H. We have proved the following corollary.

COROLLARY 5.5. *Given a basis $\{g_j\}_{j=1}^n$ for H and a dual space $P \subset \pi$ for H, there exists a unique basis $\{p_j\}_{j=1}^n$ for P satisfying for all $\theta \in \mathbb{R}^s$*

$$[0]T^{-1} = [\theta]\left(\sum_{j=1}^{n} g_j(-\theta)p_j(D)\right).$$

Moreover, $\{p_j\}_{j=1}^n$ *is the basis for* P *which is dual to* $\{f_j\}_{j=1}^n$, *where* $\{f_j\}_{j=1}^n$ *are defined by*

$$f_j \in H, \quad q(D)f_j = g_j, \qquad j = 1, \cdots, n,$$

with $q \in P$ *the unique polynomial satisfying* $q^* = [0]T^{-1}$.

For an exponential box spline ϕ there are two natural choices for a dual P for H (cf. [BDR, §4] for details).

For a *polynomial* H, we may write each of the polynomials $\{g_j\}_{j=1}^n$ in the above corollary in power form and then use summation by parts to obtain the following corollary.

COROLLARY 5.6. *Assume that* $H \subset \pi_k$ *for some nonnegative integer* k. *Then there exist polynomials* $\{p_\alpha\}_{|\alpha|\le k}$ *such that*

$$[0]T^{-1} = [\theta]\left(\sum_{|\alpha|\le k} \theta^\alpha p_\alpha(D)\right).$$

The sequence $\{p_\alpha\}_{|\alpha|\le k}$ is unique in case we impose the restriction $\{p_\alpha\} \subset P$ for a space P dual to H. We mention that in case H is invariant under the complex involution, it is self-dual and we then may choose $\{p_\alpha\} \subset H$.

Corollary 5.6 captures the construction in [J1]. There ϕ was a polynomial box spline and P was chosen as a specific known dual of H.

6. Global and local linear independence. Our preliminary objective in this note was to study possible relations between global and local linear independence. Indeed, Theorem 1.3 exhibits an equivalence between global linear independence and a very weak notion of local independence. We hoped that, at least in the translation invariant case, a stronger claim of the following type connecting global and local independence would still be valid.

CLAIM 6.1. *Assume that the integer translates of the compactly supported* $\phi \in \mathcal{D}'(\mathbb{R}^s)$ *are globally linearly independent. Then there exists a bounded set* A *that satisfies* (1.2).

Unfortunately, the above claim is true only for univariate splines. The following is a counterexample for $s = 2$, which can be lifted to higher spatial dimensions as well.

Example 6.2. Let f_1, f_2, f_3 be three univariate functions which are locally linearly independent over $[0,1]$. Define ϕ_1, ϕ_2 as functions with support on $[0,2)$ and so that

$$\phi_1 = \begin{cases} f_1 & \text{on } [0,1) \\ -f_1(\cdot - 1) & \text{on } [1,2) \end{cases}, \qquad \phi_2 = \begin{cases} f_2 & \text{on } [0,1) \\ f_3(\cdot - 1) & \text{on } [1,2) \end{cases}.$$

Construct ϕ from these two functions by

$$\phi(x,y) := \begin{cases} \phi_1(x), & 0 \le y \le 1, \\ \phi_2(x), & 1 < y < 2, \\ 0, & \text{otherwise.} \end{cases}$$

Then the integer translates of ϕ are globally linearly independent, since the vanishing of $\sum_{\alpha\in\mathbb{Z}^2} c(\alpha)\phi(\cdot-\alpha)$ on some integer square $Q_{ij} := [i, i+1]\times[j, j+1]$ implies that $c(i, j-1) = c(i-1, j-1) = 0$. So, for any given bounded A, let j be a maximal integer for which there exists an integer i such that the square Q_{ij} has some open set in common with A. Then $\sum_{k\in\mathbb{Z}} \phi(\cdot-(k, j))$ is a nontrivial linear combination of translates which vanishes on A. Defining therefore a sequence $c : \mathbb{Z}^2 \to \{0, 1\}$ by

$$c(\alpha) = \begin{cases} 1, & \alpha_2 = j \quad \text{and} \quad \operatorname{supp} E^{-\alpha}\phi \cap A \neq \emptyset, \\ 0, & \text{otherwise}, \end{cases}$$

we get $(\phi * c)_{|A} = 0$, although c is nontrivial, showing that the integer translates of ϕ are locally linearly dependent with respect to every bounded open A.

Acknowledgment. We thank Carl de Boor for his valuable suggestions made on an early draft of the paper.

REFERENCES

[B1] C. DE BOOR, *On uniform approximation by splines,* J. Approx. Theory, 1 (1968), pp. 219–235.

[B2] ———, *The polynomials in the linear span of integer translates of a compactly supported function,* Constructive Approx. 3 (1987), pp. 199–208.

[BAR] A. BEN-ARTZI AND A. RON, *Translates of exponential box splines and their related spaces,* Trans. Amer. Math. Soc., 309 (1988), pp. 683–710.

[BDR] C. DE BOOR, N. DYN, AND A. RON, *On two polynomial spaces associated with a box spline,* Pacific J. Math., to appear.

[BF] C. DE BOOR AND G. FIX, *Spline approximation by quasi-interpolants,* J. Approx. Theory, 8 (1973), pp. 19–45.

[BH] C. DE BOOR AND K. HÖLLIG, *B-splines from parallelepipeds,* J. d'Anal. Math., 42 (1982/3), pp. 99–115.

[BR1] C. DE BOOR AND A. RON, *On multivariate polynomial interpolation,* Constructive Approx., to appear.

[BR2] ———, *The exponentials in the span of the integer translates of a compactly supported function,* Comp. Sciences Tech. Report 887, University of Wisconsin-Madison, Madison, WI, 1989.

[CD] C. K. CHUI AND H. DIAMOND, *A natural formulation of quasi-interpolation by multivariate splines,* Proc. Amer. Math. Soc., 99 (1987), pp. 643–646.

[CL] C. K. CHUI AND M. J. LAI, *A multivariate analog of Marsden's Identity and a quasi-interpolation scheme,* Constructive Approx., 3 (1987), pp. 111–122.

[DM1] W. DAHMEN AND C. A. MICCHELLI, *Translates of multivariate splines,* Linear Algebra Appl., 52/3 (1983), pp. 217–234.

[DM2] ———, *On the local linear independence of translates of a box spline,* Studia Math., 82 (1985), pp. 243–263.

[DM3] ———, *Multivariate E-splines,* Advances in Math., 76 (1989), pp. 33–93.

[DR] N. DYN AND A. RON, *Local approximation by certain spaces of multivariate exponential-polynomials, approximation order of exponential box splines and related interpolation problems,* Trans. Amer. Math. Soc., 31 (1990), to appear.

[J1] R. Q. JIA, *Subspaces invariant under translation and the dual bases for box splines,* Chinese Ann. of Math., to appear.

[J2] ———, *A dual basis for the integer translates of an exponential box spline,* preprint, 1988.

[R] A. RON, *Relations between the support of a compactly supported function and the exponential-polynomials spanned by its integer translates,* Constructive Approx., 6 (1990), pp. 139–155.

[SF] G. STRANG AND G. FIX, *A Fourier analysis of the finite element variational method,* C.I.M.E. II Cilo 1971, Constructive Aspects of Functional Analysis, G. Geymonet, ed., 1973, pp. 793–840.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1563-1593, November 1990
© 1990 Society for Industrial and Applied Mathematics
011

COLLISION SINGULARITIES IN CELESTIAL MECHANICS*

MOHAMED SAMI ELBIALY†

Abstract. A general collision singularity of the n-body problem in which several clusters of particles collapse simultaneously is replaced by a collision manifold. The flow of the original problem is extended to the collision manifold. The flow on the collision manifold is studied and used to study the asymptotic behaviour of collision orbits. The work of Saari and Hulkower is generalized and it is shown that near any collision singularity of the n-body problem none of the clusters enters into an infinite spin. In order to do that, the manifold of collinear configurations is studied as an embedded submanifold of the unit sphere.

Key words. collision singularities, n-body problem, celestial mechanics, McGehee transformation, Painlevé-Wintner problem

AMS(MOS) subject classifications. 34C, 58F, 70F

1. Introduction. In this work we present a geometrical theory for studying the flow of the n-body problem near a *general collision singularity* (GCS). At a GCS several clusters of particles collapse simultaneously. This theory extends McGehee's approach which was used successfully to study the flow near triple collisions, or total collapse, in the collinear and isosceles three-body problem, [McG1], [Moc1], [Moc2]. We shall extend many of the results obtained for these problems, in which the *total collapse* is an isolated singularity, to a GCS, which is no longer isolated as we shall show below.

By studying the flow near a GCS this theory emphasizes and takes advantage of the geometrical character of collision singularities, and indeed the n-body problem in general. Instead of studying a single collision orbit, we study a collection of collision and near collision orbits. This is what makes this approach more suitable for considering questions of regularization of collision singularities, and detecting chaotic behaviour near them. It isolates and reduces the number of analytical estimates to the minimum needed by eliminating all the parts that have already been incorporated in the general theory of dynamical systems. It also isolates the parts that have a general character and handles them once and for all. As a byproduct of this approach we obtain all the essential classical results about a single collision orbit as easy corollaries with obvious geometrical meaning and with easy proofs, in sharp contrast to the lengthy and complicated asymptotic analysis arguments that were usually repeated time and again with each proof.

This work is divided into two parts, §§ 1-4 and §§ 5-7. We begin the first part by generalizing the McGehee transformation to a GCS. In order to do this, we use *cluster coordinates* to define the *singularity set* that we study. Then, we show how to introduce McGehee's coordinates in the vicinity of this *singularity set*, and how to replace it by a manifold called the *collision manifold.* The flow in these coordinates can now be extended analytically to the collision manifold which becomes invariant under this extended flow. We point out here that the *singularity set* of a total collapse is a single point while a general one is an open subset of a linear subspace of $\mathbb{R}^{3n}$ as we shall see later.

In § 3 we extend many of the results obtained for the collinear and isosceles three-body problem to this general case. We show that the flow on the collision manifold is gradientlike, and that it has a compact set of rest points for each point on the

* Received by the editors May 18, 1988; accepted for publication (in revised form) October 30, 1989.
† Institute of Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455.

singularity set. At these rest points each cluster forms a central configuration. We also show that if we pay attention to a single cluster, the quantity that is a candidate for a Lyapunov function is nondecreasing only. We end this section by giving the linearization of the vector field at a rest point, compute the eigenvalues, and show that zero eigenvalues correspond to zero eigenvalues of the Hessian of the potential when restricted to a certain ellipsoid.

In § 4 we study the asymptotic behaviour of collision orbits. We show that as in the case of collinear and isosceles three-body problem, the ω-limit set of a collision orbit is a compact subset of a single set of rest points. This is shown in Theorem 4.2, which together with a lemma by Shub [Sh] gives many of the asymptotic estimates that appear in the literature as easy corollaries (Remark 4.14). We also show that the relative size of each cluster and the radial component of the relative velocity of its particles approach definite nonzero limits. We show that this implies that no cluster collapses infinitely faster than the others. We also show that at a rest point, the flow in the direction of the relative sizes is repelling, which implies that collisions are unlikely to occur. We note that the variables giving the relative sizes (a unit vector with positive components) occur for a general collision singularity but not for a total collapse. We end this section by showing that the angular momentum of each cluster approaches zero as $t^{7/3}$, which is faster than the $t^{5/3}$ shown in [S]. We also show (Theorem 4.12) that its energy approaches a definite limit; in [Sp] it is only shown to be bounded.

In the second part, (§§ 5–7), we study the problem of *infinite spin*, also known as the *Painlevé–Wintner problem*, near a general collision singularity: can any of the clusters undergo an infinite spin as it collapses? In [S-H] it is shown that this does not happen for a total collapse. In [S] a negative answer is given for certain cases which are called "sufficiently hyperbolic." There are other cases which are not "sufficiently hyperbolic" as we will show in § 4.15(7). Accordingly, we will extend the result of [S-H] to the general case directly and without having to impose any conditions such as sufficient hyperbolicity, which are satisfied by "all currently known types of possible collision behaviour." Those types are easier to handle case by case.

In § 5, we use the action of the compact group SO (3) on $\mathbb{R}^{3m}$ to give a precise formulation of the *Painlevé–Wintner question.* We will also be able to answer this question for any cluster that stays away from collinear central configurations, or has no more than two particles. Collinear central configurations are singular with respect to the SO (3) action, since the dimension of the SO (3)-orbits drops by one when the configuration reaches a collinear one. Yet, they are nondegenerate, and hence isolated, when the flow of the m-body problem is considered. This result is due to Conley [Pc]. Moulton shows this fact when a collinear centre configuration is considered within the collinear m-body problem.

In § 6, we study the set of collinear configurations in a normalized ellipsoid Σ, show that it is an embedded submanifold $\mathbf{L}$, and give the embedding explicitly. Given $s \in \mathbf{L}$, we decompose the tangent space as $T_s\Sigma = X_s \oplus Y_s \oplus Z_s$, where X_s is the tangent to the SO (3)-orbit of s, Y_s is the subspace normal to X_s in $\mathbf{L}$, and Z_s is the subspace normal to $\mathbf{L}$. This analysis will enable us to unify the classical result of Moulton (*there are exactly $n!/2$ collinear central configurations*) and Conley's result, mentioned above, that *the Hessian of the restriction of the potential function to Σ is positive definite in the direction normal to* $\mathbf{L}$, *that is, on* $\mathbf{Z}_s$. What Moulton showed about the nondegeneracy of collinear central configurations is that this Hessian is negative definite on Y_s.

In § 7, we answer the *Painlevé–Wintner question* for collinear central configurations. This is done by showing that they are compact normal-hyperbolic invariant

manifolds for the flow of the m-body problem, and then applying the centre manifold theorem to each of them. Here m is the number of particles in the cluster under consideration.

In what follows the symbol □ will indicate the end of a proof, and when it appears at the end of a statement it will mean that the proof is by straightforward calculations.

2. Blowing up a general collision singularity. Consider the problem of n point masses moving in a three-dimensional Euclidean space under Newton's equations for gravitational force. Let $m_j > 0$ be the mass of the jth particle, and q_j its position vector. Then, the equations of motion are given by

$$M\frac{d^2q}{dt^2} = -DU(q), \tag{2.1}$$

where

$$U(q) = -\sum_{i<j}\frac{m_i m_j}{|q_i - q_j|}, \qquad M = \operatorname{diag}(m_1, m_2, \cdots, m_n),$$

$$q = (q_1, q_2, \cdots, q_n)^\dagger \in \mathbf{Q}, \qquad \mathbf{Q} = \left\{ q \in \mathbb{R}^{3n} \,\middle|\, \sum_i m_i q_i = 0 \right\} \Big/ \Delta,$$

$$\Delta = \bigcup_{i\neq j} \Delta_{ij}, \qquad \Delta_{ij} = \{q \mid q_i = q_j\},$$

and a dagger "†" stands for taking the transpose.

The right-hand side of (2.1) is C^∞ in $\mathbf{Q}$. Hence, given any set of initial conditions in TQ we can solve it for a maximal interval, say (α, σ). When $\sigma < \infty$, we say that the solution $q(t)$ *experiences a singularity as* $t \to \sigma$. In that case we assume $\sigma = 0$. In 1897, Painlevé showed that in such a case $q(t)$ approaches the diagonal set Δ as $t \to 0$ [Pn]. If $q(t)$ tends to a specific point $q^* \in \Delta$, we call the singularity a *collision singularity*, otherwise we call it a *pseudocollision*. Painlevé also showed that for $n = 3$, all singularities are collision ones. It is not yet known whether or not pseudocollisions exist for $n \geqq 4$. However, as $t \to 0$, the moment of inertia $I(t) = \frac{1}{2}q^\dagger Mq$ either diverges to infinity or goes to a finite limit. Sundman [Su] proved this assertion in 1906 for $n = 3$. In 1908, von Zeipel proved it for arbitrary n and showed that *if* $I(t)$ *is bounded as* $t \to 0$, *the singularity is actually a collision one, that is,* $q(t)$ *tends to some point* q^* *in* Δ. This is known as *von Zeipel's theorem.* (See [vZ] and [McG3].) In such a case, the n particles group themselves into several clusters, which we denote by $\mu, \cdots, \nu$, and the particles of each cluster collide as $t \to 0$. Thus, we start with a partition $\Omega = \{\mu, \cdots, \nu\}$ of the set $N = \{1, 2, \cdots, n\}$, and define the singularity set $\Delta^*(\Omega)$ associated with Ω.

DEFINITION 2.1 (The singularity set associated with a partition Ω). Let $\Omega = \{\mu, \cdots, \nu\}$ be a partition of the set $N = \{1, 2, \cdots, n\}$ and define *the singularity set associated with* Ω *by*

$$\Delta^*(\Omega) = \{q \in Q \mid q_i = q_j \text{ iff } i, j \in \mu \text{ for some } \mu \in \Omega\}. \tag{2.1.1}$$

For a total collapse, $\Omega = \Omega_0 = \{N\}$, $\Delta^*(\Omega_0) = \{0\}$. When $\Omega \neq \Omega_0$, $\Delta^*(\Omega)$ is an open subset of a linear subspace of $\mathbb{R}^{3n}$. This shows why a total collapse is an isolated singularity while a general one is not.

2.2. It is obvious that $\Delta = \bigcup_\Omega \Delta^*(\Omega)$ (disjoint union).

In order to study the singularity set $\Delta^*(\Omega)$, we need to use what are called cluster coordinates.

DEFINITION 2.3 (The cluster coordinates associated with a partition Ω). For each $\mu \in \Omega$, let $m_\mu = \sum_{i\in\mu} m_i$, and define, *the centre of mass of the cluster* μ

$$z_\mu(q) = \frac{1}{m_\mu} \sum_{i\in\mu} m_i q_i,$$

the position vector of the ith *particle in cluster* μ *relative to the cluster's centre of mass*

$$x_i(q) = q_i - z_\mu, \qquad i \in \mu,$$

and let

$$\mathbf{x}_\mu(q) = (x_i \,|\, i \in \mu), \quad \mu \in \Omega, \qquad \mathbf{x}(q) = (\mathbf{x}_\mu(q) \,|\, \mu \in \Omega),$$
$$\mathbf{z}(q) = (z_\mu(q) \,|\, \mu \in \Omega) \in \mathbb{R}^{3|\Omega|},$$

where $|\Omega|$ is the number of classes, or clusters, of Ω.

It is obvious that when $\mu \neq \nu$, $z_\mu(q^*) \neq z_\nu(q^*)$, for any $q^* \in \Delta^*(\Omega)$, and that we can find a collection of disjoint closed balls around the points $z_\mu(q^*) \in \mathbb{R}^3$, $\mu \in \Omega$.

2.4. Definitions and remarks.

(1) Define the following two sets:

$$\mathbf{Z} = \{\mathbf{z} \in \mathbb{R}^{3|\Omega|} \,|\, \mathbf{z}_\mu \neq \mathbf{z}_\nu \text{ when } \mu \neq \nu\},$$

$$\mathbf{X} = \left\{ \mathbf{x} = (\mathbf{x}_\mu | \mu \in \Omega) | \sum_{i\in\mu} m_i x_i = 0 \text{ for all } \mu \right\}.$$

Then, for each $\mu \in \Omega$, there is a continuous function $a_\mu : \mathbf{Z} \to \mathbb{R}$, $\mu \in \Omega$, such that the closures of the open balls, $B(\mathbf{z}_\mu, a_\mu(\mathbf{z}))$, $\mu \in \Omega$, are disjoint.

(2) Let

$$\mathbf{A}_z = \{(\mathbf{x}, \mathbf{z}) \,|\, \mathbf{x} \in \mathbf{X}, \|x_\mu\| < a_\mu(\mathbf{z}) \text{ for all } \mu\}, \qquad \mathbf{A} = \bigcup_{z\in Z} \mathbf{A}_z,$$

$$\mathbf{S}_0 = \{(\mathbf{x}, \mathbf{z}) \in \mathbf{A} \,|\, \mathbf{x} = \mathbf{0}\}, \qquad \mathbf{W}_\Omega = \{q \in Q \,|\, (\mathbf{x}(q), \mathbf{z}(q)) \in \mathbf{A}\}.$$

Then, $(\mathbf{x}, \mathbf{z}) : \mathbf{W}_\Omega \to \mathbf{A}$, is a diffeomorphism, and $\Delta^*(\Omega)$ is diffeomorphic to $\mathbf{S}_0$. Moreover, $\mathbf{S}_0$ itself is diffeomorphic to $\mathbf{Z}$.

(3) As a matter of fact, we shall restrict our attention to the open subset of $\mathbf{A}$ given by

$$\mathbf{E} = \bigcup_{z\in Z} \mathbf{E}_z, \qquad \mathbf{E}_z = \{(\mathbf{x}, \mathbf{z}) \in \mathbf{A}_z \,|\, x_i \neq x_j \text{ whenever } i \neq j,\ i, j \in \mu \in \Omega\}.$$

In § 4 we will show that collision orbits approach the singularity set $\mathbf{S}_0$, which is not a subset of $\mathbf{E}$, in a way that is not tangential to any of the sets $\{(\mathbf{x}, \mathbf{z}) \in \mathbf{A} \,|\, x_i = x_j\}$, for some $i \neq j$, i, j in some $\mu \in \Omega$, which justifies our choice of the open neighbourhood $\mathbf{E}$.

DEFINITION 2.5. Using the cluster coordinates we define for each cluster $\mu \in \Omega$ its *intrinsic potential energy, kinetic energy, total energy, angular momentum, and moment of inertia* respectively by,

$$U_\mu(x_\mu) = \sum_{i,j\in\mu} U_{ij}(x, z), \qquad T_\mu(y_\mu) = \frac{1}{2}\|y_\mu\|^2 = \frac{1}{2}\sum_{i\in\mu} m_i |y_i|^2,$$

$$H_\mu = T_\mu + U_\mu, \qquad c_\mu = \sum_{i\in\mu} m_i x_i \times y_i, \qquad I_\mu = \frac{1}{2}\|x_\mu\|^2,$$

where

$$y_\mu = \dot{x}_\mu,$$

$$U_{ij}(x, z) = \begin{cases} -\dfrac{m_i m_j}{2|x_i - x_j|}, & i, j \in \mu, \\ -\dfrac{m_i m_j}{2|z_\mu - z_\nu + x_i - x_j|}, & i \in \mu,\ j \in \nu,\ \mu \neq \nu. \end{cases}$$

Moreover, define the *potential energy, kinetic energy, total energy, angular momentum, and moment of inertia due to the interaction between the clusters* by

$$U_0(x, z) = \sum_{\mu \neq \nu} U_{\mu\nu}, \qquad U_{\mu\nu}(x, z) = \sum_{i \in \mu, j \in \nu} U_{ij}, \qquad \mu \neq \nu,$$

$$T_0 = \frac{1}{2}\|\pi\|^2 = \frac{1}{2}\sum_{\mu \in \Omega} m_\mu |\pi_\mu|^2, \qquad \pi = \frac{dz}{dt},$$

$$H_0 = T_0 + U_0, \qquad \mathbf{c}_0 = \sum_{\mu \in \Omega} m_\mu z_\mu \times \pi_\mu, \qquad I_0 = \frac{1}{2}\|z\|^2.$$

COROLLARY 2.6. *Let* $f_\Omega = \sum_{\mu \in \Omega} f_\mu$, *where* f *stands for* U, T, H, $\mathbf{c}$, *or* I. *Then,* $f = f_0 + f_\Omega$ *by virtue of the previous definition, and the fact that* $\sum_{i \in \mu} m_i x_i = 0$, $\mu \in \Omega$. □

2.7. Equations of motion on TE using the cluster coordinates. In the coordinate system defined above, the equations of motion take the following form where $i \in \mu \in \Omega$:

$$\dot{x}_\mu = y_\mu, \qquad \dot{y}_\mu = -M_\mu^{-1} DU_\mu(\mathbf{x}_\mu) + F_\mu, \tag{2.7.1}$$

$$\dot{z}_\mu = \pi_\mu, \qquad \dot{\pi}_\mu = G_\mu, \tag{2.7.2}$$

where

$$F_\mu = -M_\mu^{-1} D_{x_\mu} U_0(\mathbf{x}, \mathbf{z}) + \tilde{G}_\mu,$$

$$\tilde{G}_{\mu_i} = G_\mu = -\frac{1}{m_\mu} D_{z_\mu} U_0(x, z) = -\frac{1}{m_\mu} \sum_{i \in \mu} [D_{x_\mu} U_0(\mathbf{x}, \mathbf{z})],$$

$$[D_{x_\mu} U_0(\mathbf{x}, \mathbf{z})]_i = -m_i \sum_{k \in \nu, \nu \neq \mu} \frac{m_k (x_k - x_i + z_\nu - z_\mu)}{|x_k - x_i + z_\nu - z_\mu|^3},$$

$$H_\mu(x_\mu, y_\mu) = \frac{1}{2}\|y_\mu\|^2 + U_\mu(x_\mu), \qquad \mathbf{c}_\mu(x_\mu, y_\mu) = \sum_{i \in \mu} m_i x_i \times y_i,$$

$$h - H_0 = \sum_{\mu \in \Omega} H_\mu, \qquad H_0 = \frac{1}{2}\|\pi\|^2 + U_0(x, z), \qquad \sum_{i \in \mu} m_i x_i = 0,$$

$$m_\mu = \sum_{i \in \mu} m_i, \qquad M_\mu = \operatorname{diag}(m_j \mid j \in \mu), \qquad M_\Omega = \operatorname{diag}(m_\nu \mid \nu \in \Omega).$$

PROPOSITION 2.8. (a) *F and G are real analytic on* $\mathbf{E}_0 = \mathbf{E} \cup \mathbf{S}_0$.

(b) *Consider a collision singularity, where* $\mathbf{z} \to \mathbf{z}^*$ *as* $t \to 0^-$. *The velocities of the centres of mass have a finite limit as* $t \to 0^-$.

Proof. (a) The denominators of all the terms in U_0 are never zero on $\mathbf{E}_0$.

(b) Suppose the solution is defined on an interval $(-\delta, 0)$. Since $\mathbf{G}$ has a limit as $t \to 0^-$, both $\mathbf{z}$ and $\ddot{\mathbf{z}}$ have a finite limit as $t \to 0^-$. □

PROPOSITION 2.9. *The energy* H_μ, *of a cluster* $\mu \in \Omega$, *stays bounded as* $t \to 0$.

Proof. Direct computations show that $\dot{H}_\mu = \langle F_\mu, y_\mu \rangle$. Hence, we have

$$H_\mu(t) = H_\mu(t_1) + \int_{t_1}^{t} \langle F_\mu(\beta), \dot{x}_\mu(\beta) \rangle \, d\beta,$$

where $F_\mu(t) = F_\mu(x(t), z(t))$ is real analytic when $t \in [t_1, 0]$, for some $t_1 < 0$. Moreover, $x(t)$ is real analytic on $[t_1, t_0]$, for all $t_0 < 0$, and bounded on $[t_1, t_0]$. Integration by parts shows that the right-hand side is bounded. □

Proposition 2.9, which is due to Sperling [Sp], enables us to handle the situation created by the fact that the energy H_μ of the cluster μ is no longer conserved. In Theorem 4.12 we will show that each H_μ actually tends to a finite limit. Up to this point we cannot prove a similar assertion regarding the angular momentum $\mathbf{c}_\mu$. Later, we will show that $\mathbf{c}_\mu \to 0$ as fast as $t^{7/3}$, but first we must show that $\|x_\mu\| U_\mu(x_\mu)$ goes to a finite limit as x_μ tends to zero.

2.10. The McGehee transformation near a total collapse. In the case of a total collapse where the singularity set is a single point, the main steps in constructing the collision manifold and extending the flow to it are the following:

(1) Instead of the Euclidean inner product on $\mathbb{R}^{3n}$ we use the one given by the matrix M and define the corresponding polar coordinates away from the origin,

$$r = (q^\dagger M q)^{1/2}, \qquad s = r^{-1} q.$$

(2) We rescale the velocity $\dot{q}$ by a factor of $r^{1/2}$ and factorize it into the radial and tangential parts,

$$v = r^{1/2}\dot{r} \qquad w = r^{3/2}\dot{s}.$$

(3) When (2.1) is written in this coordinate system, the singularity at the origin appears as a factor of $r^{-3/2}$, which leads to the following step.

(4) If we rescale time by a factor of $r^{3/2}$, the new vector field is real analytic on $r \geqq 0$.

(5) Finally, the energy relation can be written as $2rh = v^2 + \|w\|^2 + 2U(s)$, which makes sense at $r = 0$. Actually, it gives the promised collision manifold when $r = 0$.

Remarks 2.11. (1) The first step in the McGehee transformation can be best understood if we rewrite (2.1) as $\ddot{q} = -M^{-1}DU(q)$. But $M^{-1}DU(q)$ is the gradient of U relative to the inner product given by M. Thus (2.1) becomes $\ddot{q} = -\nabla U(q)$, where ∇ denotes this gradient.

(2) The second one can be best understood if we note that the following system

$$(*) \qquad \dot{q} = y, \qquad \dot{y} = -M^{-1}DU(q),$$

is invariant under the similarity transformation $t \to \alpha^{-1}t$, $q \to \alpha^{-2/3}$, and $y \to \alpha^{1/3}$. This implies that $(*)$ has similarity solutions $q(t) = t^{2/3}q_0$, $\dot{q}(t) = \frac{2}{3}t^{-1/3}q_0$, which obviously has a collision singularity at $t = 0$. Nevertheless $r^{1/2}\dot{q} \approx \frac{2}{3}|q_0|^{3/2}$ near $t = 0$. Of course, not all collision solutions are similarity solutions, yet rescaling the velocity in this manner reflects the hope that they follow similarity solutions near the singularity and have the same limit. This actually turns out to be the case even for a general collision singularity as we will show later. Finally, going to polar coordinates, with r defined as above, is not more than blowing up the singularity set.

(3) We will show how to replace the singular set $T\mathbf{S}_0$ for a general collision singularity by a collision *manifold*, and extend the flow on $T\mathbf{E}$ to that manifold after rescaling time. In the rescaled time it takes collision solutions infinitely long to reach the singularity. We start by defining *McGehee coordinates around the singular set* $\mathbf{S}_0$. As a matter of fact there are two ways for defining what we might like to call *McGehee coordinates*, both of which are useful. The first is to apply the procedure just described to each cluster, and define a polar coordinate system around the centre of mass of that cluster, i.e., around $x_\mu = 0$. In that case we will have a vector $\mathbf{s}$ whose components are unit vectors, $\mathbf{s} = (s_\mu \mid \mu \in \Omega) \in \mathbf{S}^{|\mu|} \times \cdots \times \mathbf{S}^{|\nu|}$. The second is to apply that procedure directly to the vector $\mathbf{x} = (\mathbf{x}_\mu \mid \mu \in \Omega)$. We will use both coordinate systems.

2.12. The first McGehee-coordinate system associated with a partition Ω.

(1) Given a partition $\Omega = \{\mu, \cdots, \nu\}$ we define for each $\mu \in \Omega$ the following:

$$r_\mu = \|\mathbf{x}_\mu\|, \qquad \mathbf{s}_\mu = r_\mu^{-1}\mathbf{x}_\mu,$$
$$u_\mu = r_\mu^{1/2}\langle \dot{\mathbf{x}}_\mu, \mathbf{s}_\mu\rangle, \qquad \mathbf{w}_\mu = r_\mu^{1/2}[\dot{\mathbf{x}}_\mu - \langle \dot{\mathbf{x}}_\mu, \mathbf{s}_\mu\rangle \mathbf{s}_\mu].$$

(2) Let

$$\mathbf{r} = (r_\mu \mid \mu \in \Omega), \qquad r = |\mathbf{r}|,$$

and

$$\alpha = r^{-1}\mathbf{r}, \qquad \text{i.e.,} \quad \alpha_\mu = \frac{r_\mu}{r} > 0, \qquad \mu \in \Omega.$$

(3) Let

$$\tau(t) = \int_{t_1}^{t} r(a)^{-3/2}\, da, \qquad \text{i.e.,} \quad \frac{d\tau}{dt} = r(t)^{-3/2} > 0,$$

and let

$$K_\mu(\mathbf{s}_\mu, u_\mu, \mathbf{w}_\mu) = r_\mu H_\mu = \tfrac{1}{2}(u_\mu^2 + \|\mathbf{w}_\mu\|^2) + U_\mu(\mathbf{s}_\mu).$$

Then, *the first McGehee-coordinate system associated with the partition* Ω is given by the $(r, \alpha, \mathbf{s}, \mathbf{u}, \mathbf{w})$ and time τ where $\mathbf{s} = (\mathbf{s}_\mu \mid \mu \in \Omega)$, $\mathbf{u} = (u_\mu \mid \mu \in \Omega)$, and $\mathbf{w} = (\mathbf{w}_\mu \mid \mu \in \Omega)$.

Remarks 2.13. (1) The vector α is a unit vector whose components are all positive and represent the relative sizes of the different clusters.

(2) The vector $\mathbf{s}$ is not a unit vector, yet it belongs to the open set $\Sigma = \prod_{\mu\in\Omega} \Sigma^\mu$, where

$$\Sigma^\mu = \left\{ \mathbf{s} \in \mathbf{S}^{3|\mu|-1} \;\middle|\; s_i \neq s_j \text{ for } i \neq j,\ \sum_{i\in\mu} m_i s_i = 0 \right\} \quad \text{for all } \mu \in \Omega.$$

Furthermore, for all μ, $r^{1/2}\dot{\mathbf{x}}_\mu = u_\mu \mathbf{s}_\mu + \mathbf{w}_\mu$, and $\langle \mathbf{w}_\mu, \mathbf{s}_\mu\rangle = 0$, which means that $\mathbf{w}_\mu \in T\Sigma^\mu$ at $\mathbf{s}_\mu$.

(3) In these coordinates we have

$$\mathbf{E} = \bigcup_{z\in Z} \mathbf{E}_z, \qquad \mathbf{Z} = \{\mathbf{z} \mid z_\mu \neq z_\nu \text{ for } \mu \neq \nu\},$$

$$\mathbf{E}_z = \{(\mathbf{z}, r, \alpha, \mathbf{s}) \mid |\alpha| = 1,\ \alpha_\mu > 0,\ r\alpha_\mu \in (0, a_\mu(\mathbf{z})),\ \mathbf{s}_\mu \in \Sigma^\mu, \text{ for } \mu \in \Omega\},$$

$$\tilde{\mathbf{S}}_0 = \{(\mathbf{z}, r, \alpha, \mathbf{s}) \mid r = 0,\ \mathbf{z} \in \mathbf{Z},\ \mathbf{s}_\mu \in \Sigma^\mu,\ |\alpha| = 1,\ \alpha_\mu > 0, \text{ for } \mu \in \Omega\},$$
$$= \mathbf{Z} \times \{(r, \alpha, \mathbf{s}) \mid r = 0, |\alpha| = 1,\ \alpha_\mu > 0,\ s_\mu \in \Sigma^\mu\}$$
$$= \mathbf{Z} \times \{r = 0\} \times \Gamma \times \prod_{\mu\in\Omega} \Sigma^\mu,$$

$$\mathbf{E}_0 = \mathbf{E} \cup \tilde{\mathbf{S}}_0, \qquad \Gamma = \{\alpha \in \mathbf{S}^{|\Omega|-1} \mid \alpha_\mu > 0 \text{ for all } \mu \in \Omega\}.$$

(4) Proposition 2.9 shows that each K_μ tends to zero on collision orbits.

PROPOSITION 2.14. *In the first McGehee coordinates, the vector field on* $T\mathbf{E}$ *takes the following form, which is real analytic at* $r = 0$, *and hence can be extended to* $T\mathbf{E}_0$:

(I)
$$r' = r\sum_\nu \alpha_\nu^{1/2} u_\nu = r(\alpha^{1/2}, \mathbf{u}),$$
$$\alpha'_\mu = \alpha_\mu^{-1/2} u_\mu - (\alpha^{1/2}, \mathbf{u})\alpha_\mu,$$
$$\mathbf{s}'_\mu = \alpha_\mu^{-3/2}\mathbf{w}_\mu,$$

$$u'_\mu = \alpha_\mu^{-3/2}\left[\frac{u_\mu^2}{2} + \|\mathbf{w}_\mu\|^2 + U_\mu(\mathbf{s}_\mu) + (r\alpha_\mu)^2 f_\mu(\mathbf{x}, \mathbf{z})\right]$$

$$\mathbf{w}'_\mu = \alpha_\mu^{-3/2}\left[-\frac{u_\mu}{2}\mathbf{w}_\mu - \|\mathbf{w}_\mu\|^2 \mathbf{s}_\mu - \nabla V_\mu(\mathbf{s}_\mu) + (r\alpha_\mu)^2 \mathbf{g}_\mu(\mathbf{x}, \mathbf{z})\right],$$

$$r_\mu H_\mu = K_\mu = \frac{1}{2}(u_\mu^2 + \|\mathbf{w}_\mu\|^2) + U_\mu(\mathbf{s}_\mu),$$

where

$$V_\mu = U_\mu | \Sigma^\mu \quad \textit{and} \quad \nabla V(\mathbf{s}_\mu) = M_\mu^{-1} DU(\mathbf{s}_\mu) + U_\mu(\mathbf{s}_\mu)\mathbf{s}_\mu,$$

$f_\mu(\mathbf{x}, \mathbf{z}) = \langle \mathbf{s}_\mu, \mathbf{F}_\mu(\mathbf{x}, \mathbf{z})\rangle$ *is the component of* $\mathbf{F}_\mu$ *in the direction of* $\mathbf{s}_\mu$,

$\mathbf{g}_\mu = \mathbf{F}_\mu - f_\mu \mathbf{s}_\mu$ *is the component of* $\mathbf{F}_\mu$ *tangent to* Σ^μ,

$$K = h - H_0 = \sum_\mu H_\mu = \sum_\mu \alpha_\mu^{-1} K_\mu.$$

The equations for the $\mathbf{z}$*-variables remain unchanged.*

Proof. Straightforward calculations lead to these equations using Euler's formula which implies that $(DU(\mathbf{s}_\mu), \mathbf{s}_\mu) = -U(\mathbf{s}_\mu)$, since each U_μ is homogeneous of degree -1. Moreover, f_μ and $\mathbf{g}_\mu$ are real analytic on $\mathbf{E}_0$ by Proposition 2.8. □

DEFINITION 2.15 (*The collision manifold corresponding to a partition* Ω). The collision manifold of (I), which is a submanifold of $T\mathbf{E}_0$, is given by

$$\mathbf{C}_\Omega = T\mathbf{Z} \times \Gamma \times \mathbf{C}_0 \quad \text{where } \Gamma = \{\alpha \,|\, |\alpha| = 1,\ \alpha_\mu > 0,\ \mu \in \Omega\}$$

and

$$\mathbf{C}_0 = \left\{((\mathbf{s}_\mu, u_\mu, \mathbf{w}_\mu) \,|\, \mu \in \Omega) \,|\, (\mathbf{s}_\mu, \mathbf{w}_\mu) \in T\Sigma^\mu,\ u_\mu \in \mathbb{R},\ K = \sum_\mu \alpha_\mu^{-1} K_\mu = 0\right\}.$$

Remark 2.16. In the definition of the collision manifold we require $K = \sum_\mu \alpha_\mu^{-1} K_\mu$ to be zero. We know from Proposition 2.9 that each H_μ is bounded as r tends to zero and hence each K_μ tends to zero. Nevertheless, the energy relation $K = r(h - H_0) = rH_\Omega = \sum_\mu \alpha_\mu^{-1} K_\mu$ leads to $K = 0$. Collision solutions will be shown to tend to the subset of $\mathbf{C}_\Omega$ given by $\tilde{\mathbf{C}}_\Omega = \{p \in \mathbf{C}_\Omega \,|\, K_\mu = 0, \text{ for all } \mu \in \Omega\}$. The set $\tilde{\mathbf{C}}_\Omega$ is not necessarily a manifold. In the following proposition we will show that both $\mathbf{C}_\Omega$ and $\tilde{\mathbf{C}}_\Omega$ are invariant.

PROPOSITION 2.17. *The collision manifold is invariant under the vector field* (I). *Thus, collision orbits approach the collision manifold in infinite time* τ. *Moreover,* $\tilde{\mathbf{C}}_\Omega$ *is also invariant.*

Proof. The proposition follows immediately from the following:

$$r' = r \sum_\nu \alpha_\nu^{1/2} u_\nu,$$

$$K'_\mu = \alpha_\mu^{-3/2}[u_\mu K_\mu + (r\alpha_\mu)^2 (f_\mu + \langle \mathbf{w}_\mu, \mathbf{g}_\mu\rangle)],$$

$$K' = (\alpha^{1/2}, \mathbf{u}) K. \qquad \square$$

In Proposition 4.10 we will show that each α_μ tends to a nonzero limit on collision orbits, which will show that collision orbits approach $\tilde{\mathbf{C}}_\Omega$.

PROPOSITION 2.18. *The vector field* (I) *has no rest points outside the collision manifold.*

Proof. The time scaling and change of variables that we applied are all diffeomorphisms outside $\mathbf{C}_\Omega$. If (I) had a rest point outside $\mathbf{C}_\Omega$, it would correspond to a rest point for the original system. This cannot happen, for if m_j is a particle of maximum distance from the centre of mass (the origin), and P is the plane, in $\mathbb{R}^3$, whose normal is q_j, then all the other particles would lie on one side of P, namely, the side of the centre of mass. In such a configuration, the nonzero forces that these particles exert on m_j cannot balance each other to give a zero resultant. Hence there are no rest points outside $\mathbf{C}_\Omega$. □

Remarks 2.19. (1) As we promised in the beginning of this section, we have shown how to replace the singularity at $T\mathbf{S}_0$ by an invariant manifold $\mathbf{C}_\Omega$, and to extend the flow to that manifold. We have also shown that collision orbits are slowed down in the rescaled time so that they approach $\mathbf{C}_\Omega$ in infinite τ-time. In Proposition 3.4 we will show that the extended vector field has rest points on $\mathbf{C}_\Omega$ while the original one did not have any.

(2) $\mathbf{E}_0$ is not compact since it was constructed by choosing a small neighbourhood of each q^* in $\Delta^*(\Omega)$. Moreover, these neighbourhoods cannot be chosen uniformly. As a matter of fact, their diameters shrink to zero as q^* approaches $\Delta(\Omega')$ for any Ω' obtained from Ω by combining some of the clusters of Ω together. Physically, this means that the clusters might be arbitrarily close to each other at the moment of simultaneous collapse. Nevertheless, $\mathbf{E}_0$ is locally compact and we will take advantage of that and of von Zeipel's theorem that we mentioned in the Introduction, when we study the asymptotic behaviour of collision orbits.

(3) In § 2.11 above, we defined McGehee coordinates for each cluster separately and then defined the vector α, which couples the equations of all the clusters together and whose components give the relative sizes of different clusters. In the next article we define McGehee-coordinates for the system as a whole because some of the assertions which we will make later will be easier to prove when the vector field is written in these coordinates.

2.20. The second McGehee-coordinate system associated with a partition Ω. Given a partition Ω, we define the following:

$$r = \|\mathbf{x}\| = \sqrt{\sum_\mu r_\mu^2}, \qquad \sigma = r^{-1/2}\mathbf{x},$$

$$v = r^{1/2}(\sigma, \dot{\mathbf{x}}) = \sum_\nu \alpha_\nu^{1/2} u_\nu, \qquad \gamma = r^{1/2}\dot{\mathbf{x}} - v\sigma,$$

$$\frac{d}{d\tau} = r^{3/2}\frac{d}{dt}.$$

Straightforward calculations show that $v = \sum_\nu \alpha_\nu^{1/2} u_\nu$.

PROPOSITION 2.21. *In these variables, the vector field* (2.7.1) *takes the following form, which is real analytic on* $\mathbf{E}_0$:

(II)
$$r' = rv = \sum_\nu \alpha_\nu^{1/2} u_\nu,$$

$$\sigma' = \gamma,$$

$$v' = \frac{1}{2}v^2 + \|\gamma\|^2 + V(\sigma) + r^2 g = \frac{1}{2}\|\gamma\|^2 + K + r^2 g,$$

$$\gamma' = -\frac{v}{2}\gamma - \|\gamma\|^2\sigma - \nabla V(\sigma) + r^2\mathbf{G},$$

where

$$K = r(h - H_0) = \tfrac{1}{2}(v^2 + \|\gamma\|^2) + V(\sigma),$$
$$g = \langle \mathbf{F}, \sigma \rangle, \qquad \mathbf{G} = \mathbf{F} - g\sigma, \qquad V = U_\Omega | \Sigma,$$
$$\Sigma = \left\{ \mathbf{x} \,|\, \|\mathbf{x}\| = 1,\ x_i \neq x_j \text{ for } i \neq j,\ i, j \in \mu \in \Omega;\ \sum_i m_i \mathbf{x}_i = \mathbf{0} \right\},$$

and a dash denotes differentiation with respect to τ. □

3. The flow on the collision manifold. We shall devote this section to studying the flow on the collision manifold. The collision manifold by itself does not have any physical meaning. Nevertheless, some of the properties of the flow on it are useful in describing the asymptotic behaviour of collision orbits as well as studying near collision orbits. First, we shall give the vector field on the collision manifold explicitly. It will turn out that this flow has a Lyapunov function. We shall describe the set of rest points of that vector field on $\mathbf{C}_\Omega$ which has already been shown in Proposition 2.18 to be the set of rest points for the whole system. Also, we shall give the linearization of the vector field near rest points.

3.1. The vector field on the collision manifold. The vector field on $\mathbf{C}_\Omega$ is given in the first McGehee coordinates by

$$z'' = 0,$$

and

(I_0)
$$\alpha'_\mu = \alpha_\mu^{-1/2} u_\mu - (\alpha^{1/2}, \mathbf{u})\alpha_\mu,$$
$$\mathbf{s}'_\mu = \alpha_\mu^{-3/2} \mathbf{w}_\mu,$$
$$u'_\mu = \alpha_\mu^{-3/2}\left[\frac{u_\mu^2}{2} + \|\mathbf{w}_\mu\|^2 + U_\mu(\mathbf{s}_\mu)\right] = \alpha_\mu^{-3/2}\left[\frac{\|\mathbf{w}_\mu\|^2}{2} + K_\mu\right]$$
$$\mathbf{w}'_\mu = \alpha_\mu^{-3/2}\left[-\frac{u_\mu}{2}\mathbf{w}_\mu - \|\mathbf{w}_\mu\|^2 \mathbf{s}_\mu - \nabla V_\mu(\mathbf{s}_\mu)\right],$$
$$K_\mu = \frac{1}{2}(u_\mu^2 + \|\mathbf{w}_\mu\|^2) + U_\mu(\mathbf{s}_\mu), \qquad \sum_\mu K_\mu = 0.$$

In the second McGehee coordinates, (I_0) takes the form

(II_0)
$$\sigma' = \gamma,$$
$$v' = \frac{1}{2} v^2 + \|\gamma\|^2 + V(\sigma) = \frac{1}{2}\|\gamma\|^2,$$
$$\gamma' = -\frac{v}{2}\gamma - \|\gamma\|^2 \sigma - \nabla V(\sigma),$$

where $K = \frac{1}{2}(v^2 + \|\gamma\|^2) + V(\sigma) = 0$.

DEFINITION 3.2. A flow on a manifold M is said to be *gradientlike with respect to a function* L if and only if L is strictly increasing except on rest points. L is called a *Lyapunov function.*

COROLLARY 3.3. *The flow on the collision manifold is gradientlike with respect to* v.

Proof. On $\mathbf{C}_\Omega$, $v'(\tau) = \frac{1}{2}\|\gamma\|^2 \geqq 0$. Moreover, if $v'(\tau) = 0$, $\gamma(\tau) = 0$, and hence, $\sigma'(\tau) = 0$. If $\gamma'(\tau) = 0$, we have a rest point. If $\gamma' \neq 0$, γ will vary and hence v continues to increase. □

PROPOSITION 3.4 (The rest points). (a) *A point* $((0, \alpha_\mu, \mathbf{s}_\mu, u_\mu, \mathbf{w}_\mu)^* \mid \mu \in \Omega) \in \mathbf{C}_0$ *is a rest point of* (I_0), *and* (II_0) *if and only if for all* $\mu \in \Omega$,

$$\text{(i)}\quad \mathbf{w}_\mu^* = \mathbf{0}, \qquad \text{(ii)}\quad K_\mu = 0,$$

$$\text{(3.4.1)}\qquad \text{(iii)}\quad (u_\mu^*)^2 + 2U_\mu(\mathbf{s}_\mu^*) = 0, \qquad \text{(iv)}\quad \nabla V_\mu(\mathbf{s}_\mu^*) = \mathbf{0},$$

$$\text{(v)}\quad u_\mu^* = (\alpha_\mu^*)^{3/2}(\alpha^{*1/2}, \mathbf{u}^*) = (\alpha_\mu^*)^{3/2} v^*.$$

(b) *Since* U_μ *never vanishes,* $u_\mu^* \neq 0$, *for all* μ. *Moreover, from* (v) *above, it follows that the* $\operatorname{sign}(u_\mu^*) = \operatorname{sign}((\alpha^{*1/2}, \mathbf{u}^*)) = \operatorname{sign}(v^*)$ *for all* μ *at a rest point.*

(c) *For such a point, each* α_μ^* *is uniquely determined by being positive and by*

$$\text{(3.4.2)}\qquad \alpha_\mu^* = \frac{\lambda_\mu}{|\lambda|}, \qquad \lambda_\mu = |u_\mu^*|^{2/3} > 0, \qquad \lambda = (\lambda_\mu \mid \mu \in \Omega),$$

and hence, each α_μ^* *is bounded away from zero and one.* □

Remarks 3.5. (1) The rest points are actually on $\tilde{\mathbf{C}}_\Omega$ by virtue of (ii).

(2) At a rest point $\mathbf{s}_\mu^*$ is a critical point of the restriction V_μ of the potential function of the cluster μ to the part of the unit sphere where it is defined. The set of critical points of V_μ plays an important role in determining the behaviour of the cluster μ near collisions. The structure of this set is not very simple, and we shall devote the better part of §§ 5–7 to studying parts of this set. For the moment we observe that if $\mathbf{s}_\mu^*$ is such a critical point, and A is in SO (3), then $A\mathbf{s}_\mu^*$ is a rest point as well, where $(A\mathbf{s}_\mu)_i = A\mathbf{s}_i$ for $i \in \mu$. This is true because V_μ is invariant under the action of SO (3). Moreover, the SO (3) orbits of critical points need not be isolated. In fact, if the Hessian of V_μ is nondegenerate in the direction normal to the SO (3) orbit of $\mathbf{s}_\mu^*$, then this orbit is isolated. We will see that this is the case if $\mathbf{s}_\mu^*$ is collinear, that is, if all the particles of μ lie on the same line. For the moment, we will state a lemma about these critical points which is due to Shub [Sh].

LEMMA 3.6 (Shub [Sh, 1971]). *The critical points of* V_μ *lie in a compact subset of* Σ^μ, *and hence are compact.* □

COROLLARY 3.7. *The set of rest points of* (I) *for a fixed value of* $(\mathbf{z}, \dot{\mathbf{z}})$ *is compact.*

Proof. The set of rest points is closed because it is the zero set of a continuous function. Moreover, all the variables at a rest point are bounded except, perhaps, u_μ. But $u_\mu^2 = -2V_\mu(\mathbf{s}_\mu)$, and Shub's lemma implies that V_μ is finite at a critical point. □

DEFINITION 3.8 (Definition of central configurations). Let $\mathbf{s}_\mu$ be a critical point of V_μ. The SO (3) orbit of $\mathbf{s}_\mu$ is $\{A\mathbf{s}_\mu \mid A \in \mathrm{SO}(3)\}$, and is called *a central configuration.* We shall denote it by $\mathrm{Orb}(\mathbf{s}_\mu)$.

Remarks 3.9. (1) Central configurations correspond to the so-called similarity solutions as follows. If we have only one cluster with potential U, and if $\mathbf{s}$ is a critical point of the restriction $V = U \mid \Sigma$, then *similarity solutions* are of the form $x(t) = a(t - t^*)^{2/3}\mathbf{s}$. Obviously, these solutions do not change their configurations, and this is why they are called *similarity solutions.* It is also clear that they are collision solutions and that $\dot{x}$ becomes unbounded as $t \to t^*$. In the next section we will show that, although not all collision solutions are similarity solutions, they approach the set of similarity solutions asymptotically.

(2) The vector α has norm one, and its components represent the relative sizes of the different clusters. Later we shall see that collision orbits approach the set of rest points, and we have already seen that the α_μ's are bounded away from zero on rest points. This will mean that no cluster collapses faster than the others.

(3) Although $u_\mu' \geqq 0$, u_μ is not a Lyapunov function for the motion of the cluster μ. In order to see this we fix a subset $\Gamma(y^*) = \Gamma \times \{y^*\} \subset \tilde{\mathbf{C}}_\Omega$, where $y^* = (r, \mathbf{s}, \mathbf{u}, \mathbf{w})^*$

and (α^*, y^*) is a rest point with $u_\mu^* < 0$ for all $\mu \in \Omega$. In other words, we fix all variables, except α, at a rest point with negative u_μ's. Since α does not appear in the energy relation $K = 0$, α is not forced to be α^*, and we have the following proposition.

PROPOSITION 3.10. *Let $\Gamma(y^*)$ be as above and note that it is a regular submanifold diffeomorphic to Γ. Let $g(\alpha) = 2(b, \alpha^{1/2}) = -2v > 0$, where $b_\mu = -u_\mu^* > 0$.*

(a) *$\Gamma(y^*)$ is invariant, and the flow on it is a gradient one given by*

$$\alpha' = -\nabla g(\alpha). \tag{1}$$

(b) *This flow has a unique fixed point at $p^* = (\alpha^*, y^*)$.*

(c) *The eigenvalues of the linearized flow are all positive, and hence p^* is a repeller.*

(d) *Moreover, since equation (1) is exactly the one for finding the extreme points of g, $g(\alpha^*)$ is maximum.*

(e) *When $u_\mu^* > 0$ for all μ, we have a similar flow in which all the eigenvalues at p^* are negative.*

Proof. (a) First we observe that

$$\nabla g(\alpha) = Dg(\alpha) - (Dg(\alpha), \alpha)\alpha = Dg(\alpha) - \tfrac{1}{2} g(\alpha)\alpha. \tag{2}$$

But,

$$\alpha_\mu' = -\frac{b_\mu}{\sqrt{\alpha_\mu}} + (\alpha^{1/2}, \mathbf{b})\alpha_\mu. \tag{3}$$

(b) The above system has a unique fixed point at $\alpha = \alpha^*$, which was completely determined in terms of the vector $\mathbf{b}$ in Proposition 3.4. More precisely, $\lambda_\mu = b_\mu^{2/3}$.

(c) In order to show that the eigenvalues are all positive, we compute the Hessian of g at α^*, and show that it is negative definite. From (2), it follows that if $X \in T_{\alpha^*}\Gamma$, i.e., $(\alpha^*, X) = 0$, we have

$$\begin{aligned} Hg(\alpha^*)(X, X) &= \left(\frac{d}{dt}\right)_{t=0} (X, \nabla g(\alpha^* + tX)) \\ &= D^2 g(\alpha^*)(X, X) - \frac{1}{2}(X, \alpha^*)(Dg(\alpha^*), X) - \frac{1}{2} g(\alpha^*)|X|^2 \\ &= -\sum_\mu \frac{1}{2}[b_\mu (\alpha_\mu^*)^{-3/2} + g(\alpha^*)]X_\mu^2, \end{aligned}$$

as

$$(\alpha^*, X) = 0, \qquad \frac{\partial^2 g}{\partial \alpha_\mu \, \partial \alpha_\nu} = \delta_{\mu\nu} \frac{\partial^2 g}{\partial \alpha_\mu^2} = -\delta_{\mu\nu} \frac{b_\mu}{2[\alpha_\mu^*]^{-3/2}}.$$

Parts (d) and (e) are obvious. □

Remark. In the previous proposition we considered only the case when all the u_μ's are negative because we have seen in Proposition 3.4 that at rest points all the u_μ's have the same sign. Moreover, we shall show in the next section that on a collision orbit each u_μ approaches a negative limit. Similarly, on an ejection orbit they all approach positive values as $\tau \to -\infty$. For the time being, we return to the flow on the collision manifold.

3.11. The linearization of the vector field at a rest point. Straightforward calculations show that the linearization of the vector field (I) near a rest point

$(0, \alpha, \mathbf{u}, \cdots, s_\mu, w_\mu, \cdots)^*$ is given by the matrix

$$\begin{bmatrix} v & 0 & 0 & 0 \\ 0 & -J & \# & 0 \\ 0 & 0 & W & 0 \\ 0 & 0 & 0 & B \end{bmatrix}^*, \qquad B = \operatorname{diag}(B_\mu \mid \mu \in \Omega),$$

$$B_\mu = (\alpha_\mu^*)^{-3/2} \begin{pmatrix} 0 & I \\ -J & a_\mu I \end{pmatrix}^*, \qquad a_\mu = -\frac{u_\mu^*}{2}, \quad \mu \in \Omega,$$

where

$$v = v^* = (\alpha^{*1/2}, \mathbf{u}^*), \qquad W = \operatorname{diag}((u_\mu / \alpha_\mu^{3/2})^* \mid \mu \in \Omega),$$

and

$$J = Hg(\alpha^*), \qquad J_\mu = HV_\mu(S_\mu^*),$$

and I is the $3|\mu| \times 3|\mu|$ identity matrix. We have omitted the part corresponding to the variable z, for it consists of r^2 times a finite term, and hence vanishes on the collision manifold.

PROPOSITION 3.12. *The eigenvalues of the above matrix are*: v^*; *the eigenvalues of* $-J$ *which were shown in Proposition* 3.10 *to be all positive* (*if* $v^* < 0$) *or all negative* (*if* $v^* > 0$); *the eigenvalues of* W, *namely*, $(u_\mu \alpha_\mu^{-3/2})^*$, *for* $\mu \in \Omega$; *and*

$$\kappa_\mu^{\pm} = \tfrac{1}{2}(\alpha_\mu^*)^{-3/2}(a_\mu \pm \sqrt{a_\mu^2 - 4\lambda_\mu}), \qquad \mu \in \Omega, \tag{3.12.1}$$

where λ_μ *is an eigenvalue of the Hessian of* V_μ.

Proof. Only the last set of eigenvalues needs an explanation. Let y_μ be an eigenvector of J_μ with λ_μ as an eigenvalue. Then, $(y_\mu, cy_\mu)^\dagger$ is an eigenvector $\alpha_\mu^{3/2} B_\mu$ with eigenvalue c if and only if $[J_\mu - (a_\mu c - c^2)I] = 0$, which occurs if and only if $c = \alpha_\mu^{3/2} \kappa_\mu$. □

The previous proposition shows that the critical points of V_μ determine the properties of the flow close to the set of rest points. In the following section, we shall show that collision orbits approach the set of rest points. Since the other eigenvalues do not qualitatively depend on the masses it follows that actually the critical points of V_μ determine the asymptotic behaviour of the cluster μ on a collision orbit.

4. The asymptotic behaviour of a collision orbit. We shall devote this section to studying the asymptotic behaviour of a collision orbit. The main theorem of this section is Theorem 4.2, which states that a collision orbit approaches the set of critical points on the collision manifold. Recall that on a collision singularity $(\mathbf{z}, \dot{\mathbf{z}})$ has a finite limit $(\mathbf{z}, \dot{\mathbf{z}})^*$ (Proposition 2.8), and that over each $(\mathbf{z}, \dot{\mathbf{z}})$ there is a compact set of rest points (Proposition 3.4).

DEFINITION 4.1. Let ϕ be the flow generated by the vector field given by (I), or equivalently by (II), of the previous section. We define the *ω-limit set of a subset* Y of $\mathbf{E}$ to be

$$\omega(Y) = \bigcap_{\tau > 0} \overline{\phi(Y, [\tau, \infty[)}.$$

We remark that an ω-limit set is invariant, and if it is a subset of a compact set, it is nonempty, closed, and connected whenever Y is (see [CC] and [McG4]). The ω-limit set of a collision orbit is obtained by letting $Y = \{p\}$ for some point p on it. Let $\phi(p, \tau)$, $\tau \geqq 0$, be such a collision orbit for some point $p \in \mathbf{E}$, with $r(p)$ small. In what follows we fix p, with $(z, \dot{z}) \to (z, \dot{z})^*$, and write $\phi(\tau)$ instead of $\phi(p, \tau)$ for

simplicity. Now we state the main theorem in this section and two lemmas that are needed to prove it.

THEOREM 4.2. *The ω-limit set of a collision orbit is a subset of the compact set of rest points over $(\mathbf{z}, \dot{\mathbf{z}})^*$.*

LEMMA 4.3. *As $r \to 0$ and $\tau \to \infty$, v tends to a negative limit $v^* < 0$.*

LEMMA 4.4. *The ω-limit set of a collision orbit is not empty.*

4.5. Proof of Lemma 4.3. (i) First we show that $v(\tau)$ is bounded above by a negative number. Recall the vector field (II) of Proposition 2.21. If $v(\tau)$ was always positive for large τ, r would not go to zero. Thus, $v(\tau)$ must be negative on some increasing sequence $\tau_n \to \infty$. Moreover, on the set given by $v = 0$, we have

$$v' = (-v^2/2 - V(\sigma) + 2K + r^2 g(\tau))_{v=0} = -V(\sigma) + 2K + r^2 g(\tau).$$

But $(-V(\sigma))$ has a positive minimum, and the rest of the right-hand side tends to zero as $r \to 0$. Thus, there are constants $r_- > 0$, $v_- < 0$, and $b > 0$, such that, for $|v| \leqq |v_-|$ and for $r \leqq r_-$, we have $v'|_{v=0} > b > 0$. Now define the following block B and its sides A_-, B_0, and B_- as follows.

$$\begin{aligned} &B = \{x \in \mathbf{E} \mid r \leqq r_-,\ v_- \leqq v \leqq 0\}, && A_- = \{x \in B \mid r = r_-\}, \\ &B_0 = \{x \in B \mid v = 0\}, && B_- = \{x \in B \mid v = v_-\}. \end{aligned} \tag{4.5.0}$$

Thus, $v' > b$ inside the block B, and hence, if a solution passes through B_-, it must cross B_0 in finite time. Moreover, $r' < 0$ on the side A_-, that is, the vector field is transversal to A_- and points inside B. See Fig. 4.1 for the illustration. On a collision solution $r \to 0$ as $\tau \to \infty$, and hence, for some τ_1, $r(\tau) < r_-$ for all $\tau > \tau_1$. If this solution entered B, v would have to become positive in finite time and would have to stay positive, for the vector field on B_0 points in the positive direction of v, and r is less than r_- for large τ. This would force r to increase and not go to zero, a contradiction.

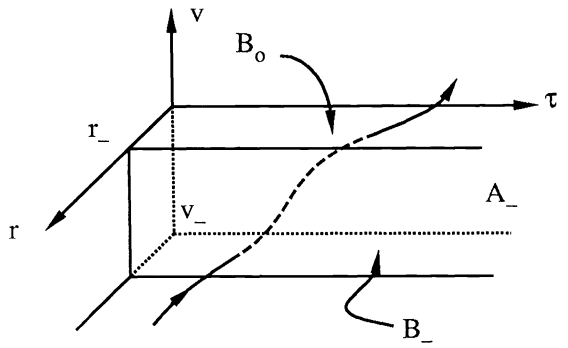


FIG. 4.1

(ii) Second, we show that v actually tends to a limit. Note that v is negative for large τ, and hence r is decreasing. From the vector field (II) we obtain that for large τ,

$$(\ln r)' = v < v_- < 0, \tag{4.5.1}$$

and hence,

$$r(\tau) \leqq r(\tau_0) \exp[v_-(\tau - \tau_0)], \qquad \tau > \tau_0. \tag{4.5.2}$$

Thus,

$$v' = \frac{\|\gamma\|^2}{2} + K + r^2 g(\tau) \geqq r(\tau)[h - H_0 + rg] > -Br(\tau), \tag{1}$$

for some positive constant B. The last inequality holds because both $(h-H_0)$ and g are bounded as $r\to 0$. Hence, using (4.5.2), we get

$$v(\tau)-v(\tau_0)>\frac{Br_0}{v_-}[1-\exp(v_-(\tau-\tau_0))] \quad \text{for } \tau>\tau_0. \tag{2}$$

Note that by choosing large τ_0, $r_0=r(\tau_0)$ can be made arbitrary small. Since v_- is negative, the right-hand side of (2) can be made bigger than any negative number which is arbitrary close to zero. Now, let $v_0=\limsup_{\tau\to\infty} v(\tau)<v_-<0$. For $\varepsilon>0$, choose τ_0 large enough so that, $v(\tau)<v_0+\varepsilon$, $0<v_0-v(\tau_0)<\varepsilon$, and $v(\tau)-v(\tau_0)>-\varepsilon$ for $\tau>\tau_0$. Thus, $-\varepsilon<v_0-v(\tau)<2\varepsilon$, for $\tau>\tau_0$. □

COROLLARY 4.6. *Since v goes to a negative limit, there is a lower bound on the decay of r in addition to the inequality* (4.5.2). *Thus we have, as $t\to 0$ and $\tau\to\infty$,*

$$b_1\exp(v_1\tau)<r(\tau)<b_2\exp(v_2\tau) \quad \textit{for some } v_1, v_2<0, \textit{ and } b_1, b_2>0. \tag{4.6.1}$$

Moreover, since $d\tau/dt=r^{-3/2}$, it follows that $r(t)\approx t^{2/3}$. □

4.7. Proof of Lemma 4.4. We want to show that $\omega(p)\neq\varnothing$, where p is any point on a collision orbit. First we show that there is an increasing sequence $\tau_n\to\infty$, such that $\gamma(\phi(\tau_n))\to 0$, as $n\to\infty$. In order to do that let $\beta(\tau)=\|\gamma(\phi(\tau))\|$ and *assume* that *there is $a>0$, and $\tau_1>0$ such that $\beta(\tau)>2a$ for all $\tau>\tau_1$*. We can take τ_1 large enough so that $K+r^2f>-a^2$, and hence $v'>a^2$ for all $\tau>\tau_1$. Thus, $v(\tau)>v(\tau_1)+a^2(\tau-\tau_1)$, which goes to infinity as $\tau\to\infty$, contradicting Lemma 4.3. Thus, our assumption is wrong and we are left with its converse: *for all $a>0$, and for all $\tau_1>0$, there is $\tau_2>\tau_1$ such that $\beta(\tau_2)\leqq a$*. It follows that *there is an increasing sequence $\tau_n\to\infty$ such that, $\gamma(\tau_n)\to 0$ as $n\to\infty$*. Since v and σ are bounded, it follows, perhaps after passing to a subsequence, that the sequence $(\phi(p,\tau_n)\,|\,n=1,2,3,\cdots)$ converges to a point on the collision manifold which shows that $\omega(p)\neq\varnothing$. □

4.8. Proof of Theorem 4.2. Let $p_0\in\omega(p)$ and let $\tau_n\to\infty$ be an increasing sequence such that $\phi(p,\tau_n)\to p_0$ as $n\to\infty$, and let $v^*=\lim_{\tau\to\infty} v(\phi(p,\tau))$. Assume that p_0 is not a rest point. Let $\varepsilon>0$ be small and let $q=\phi(p_0,\varepsilon)$. Since v' is positive on the collision manifold except at rest points, it follows that $v^*=v(p_0)<v(q)$. Then, $q=\phi(p_0,\varepsilon)=\lim_{n\to\infty}\phi(p,\tau_n+\varepsilon)$ since ϕ is a continuous flow. Hence, $v(q)=\lim_{n\to\infty} v(\phi(p,\tau_n+\varepsilon))=v^*<v(q)$, which is absurd. Thus, p_0 is a rest point. □

COROLLARY 4.9. *$\omega(p)$ is compact, for it is closed and is contained in a compact subset of rest points as was shown in Corollary* 3.7. □

In the last few paragraphs we viewed all the clusters as one system. In the next few we shall study the behaviour of each single cluster as they approach simultaneous collapse. Theorem 4.2 above and Proposition 3.4 imply that, as $r\to 0$, the variables of each cluster μ approach the set of rest points.

PROPOSITION 4.10. *On a collision orbit, as $\tau\to\infty$, and $r\to 0$, the following is true*:

(a) *For large τ, each u_μ is bounded above by a negative number, and since they are finitely many they have a common bound $u_0<0$.*

(b) *There are numbers $0<a<b<1$ such that, for large τ, and for all μ, $a<\alpha_\mu(\tau)<b$.*

(c) *Actually, each u_μ tends to a negative limit $u_\mu\to u_\mu^*<0$.*

(d) *For each μ, $\alpha_\mu\to\alpha_\mu^*=(u_\mu^*/v^*)^{2/3}$.*

(e) *Each $\mathbf{s}_\mu$ tends to the set $\Sigma^{\mu^*}=\{\mathbf{s}_\mu^*\in\Sigma^\mu\,|\,\nabla V_\mu(\mathbf{s}_\mu^*)=0\}$.*

(f) *In fact, $\mathbf{s}_\mu$ tends to a level set of V_μ in Σ^{μ^*}. That is, regardless of whether $\mathbf{s}_\mu$ has a limit or not $\nabla V_\mu\to 0$ and $V_\mu(\mathbf{s}_\mu)\to-(u_\mu^*)^2/2$, as $\tau\to\infty$.*

Proof. (a) The proof of this part is similar to that of § 4.5(i) above.

(b) As $\tau \to \infty$, v has a finite limit and u_μ is bounded away from zero. Moreover, $\alpha'_\mu = 0$ if and only if $u_\mu = \alpha_\mu^{3/2} v$. Thus, α'_μ cannot vanish if α_μ is less than a small positive δ. If α_μ is not bounded away from zero, there is large τ_0, such that $\alpha_\mu(\tau_0) < \delta$, and $\alpha'_\mu(\tau_0)$ is less than $(-k)$, for arbitrary large k. But in that case, $\alpha'_\mu(\tau)$ will continue to be negative for $\tau > \tau_0$. Thus, α_μ will continue to decrease which makes $\alpha'_\mu(\tau) < -k$ for all $\tau > \tau_0$. This will force α_μ to become negative in finite time, contradicting the fact that it is always nonnegative. Thus, each α_μ is bounded away from zero, and since they are finitely many, a bound can be chosen uniformly. Also, since $|\alpha| = 1$, it follows that they are bounded away from one.

(c) Given (b) above, we can prove that u_μ goes to a limit as we did for v in § 4.5(ii) above. Furthermore, knowing that both K_μ and w_μ go to zero, we see that the limit takes the value claimed above.

Parts (d) and (e) follow from Proposition 3.4 and Theorem 4.2.

(f) $V_\mu(\mathbf{s}_\mu)$ tends to $-(u_\mu^*)^2/2$, since both w_μ and K_μ go to zero. □

In the previous proposition we had to divide the assertion about u_μ into two parts because we cannot prove that u_μ actually goes to a negative limit without knowing that α_μ is bounded away from zero (for the vector field (I) has a factor of $\alpha_\mu^{-3/2}$ in the u_μ-component). But to prove this assertion about α_μ, we need to know that u_μ is bounded away from zero.

We shall use the last proposition to prove a slight generalization of a theorem of Sundman [Su] about the angular momentum of each cluster μ, which is no longer a constant of motion and need not be identically zero in order for a collision to take place.

PROPOSITION 4.11 (a slight generalization of a theorem by Sundman). *On a collision orbit, as $r \to 0$, $t \to 0$, and $\tau \to \infty$, the intrinsic angular momentum $\mathbf{c}_\mu$ of the cluster μ tends to zero and satisfies*

$$|\mathbf{c}_\mu(\tau)| < b r^{7/2} \approx b \exp[(7/2) v_- \tau] \approx b t^{7/3}, \tag{4.11.0}$$

where $v_- < 0$, and b stands for a positive bounded constant that might vary from one expression to another.

Proof. First we show that $\mathbf{c}_\mu \to 0$. Let $\mathbf{y} = \dot{\mathbf{x}}$. Since $\mathbf{c}_\mu = \sum_{i \in \mu} m_i (x_i \times y_i)$, it follows that

$$|\mathbf{c}_\mu|^2 \leqq \|\mathbf{x}_\mu\|^2 \|\mathbf{y}_\mu\|^2 \leqq 2 r_\mu^2 T_\mu(t) = 2 r_\mu^2 [H_\mu - r_\mu^{-1} V_\mu(\mathbf{s}_\mu)]. \tag{4.11.1}$$

Although $\mathbf{s}_\mu$ might not have a limit, it tends to the set of critical points of V_μ. This set is compact by Shub's Lemma 3.6 [Sh]. Hence, V_μ stays bounded as $r \to 0$. We have also seen that H_μ is bounded. Thus, $\mathbf{c}_\mu \to 0$. Now consider the following expression:

$$\dot{\mathbf{c}}_\mu = \sum_{i \in \mu} m_i (x_i \times \dot{y}_i) = -\sum_{i \in \mu} x_i \times D_i U_\mu(\mathbf{x}_\mu) + \sum_{i \in \mu} m_i (x_i \times F_i). \tag{4.11.2}$$

The first sum on the right-hand side vanishes because the terms involving $x_i \times x_j$ and $x_j \times x_i$ cancel out. Moreover, each F_i is a linear combination of terms of the form

$$\frac{a - x_i}{|a - x_i|^3} = |a|^{-3}(1 + r b_i)(a - x_i), \quad a = z_\nu + x_k - z_\mu, \quad k \in \nu \neq \mu. \tag{4.11.3}$$

Each b_i is bounded, and the coefficients of the linear combination are constants and do not depend on i. Since $\sum_{i \in \mu} m_i x_i = 0$, it follows that $\dot{\mathbf{c}}_\mu = r^2 \mathbf{b}(r(t))$, where $\mathbf{b}$ stands for the function bounded as $r \to 0$. From the relation between τ and t it follows that $\mathbf{c}'_\mu = r^{7/2} \mathbf{b}(\tau)$. For large τ, $r' = rv < 0$ since v has a negative limit $v^* < 0$. Thus, $|d\mathbf{c}/dr| < b r^{5/2}$, and we use the mean value theorem to get

$$|\mathbf{c}_\mu(r) - \mathbf{c}_\mu(r_0)| < b r^{5/2}(r - r_0), \qquad r_0 < r. \tag{4.11.4}$$

If we allow r_0 to go to zero, $\mathbf{c}_\mu(r_0)$ will also go to zero. Thus, $|\mathbf{c}_\mu(r)| < b r^{7/2} = b t^{7/3}$ by (4.6.1). □

THEOREM 4.12. *On collision orbits each H_μ tends to a finite limit as fast as r.*

Proof. Recall from Proposition 2.17 that $H_\mu = r_\mu K_\mu$, and that

$$K'_\mu = \alpha_\mu^{-3/2}[u_\mu K_\mu + (r\alpha_\mu)^2(f_\mu + \langle \mathbf{w}_\mu, \mathbf{g}_\mu \rangle)].$$

Moreover, $r'_\mu = r_\mu u_\mu$ and $u_\mu \to u^*_\mu < 0$. Thus,

$$H'_\mu = \frac{r_\mu}{\alpha_\mu^{3/2}}[f_\mu + \langle \mathbf{w}_\mu, \mathbf{g}_\mu \rangle].$$

But α_μ tends to $\alpha^*_\mu > 0$ (Proposition 4.10(d)) and the square bracket is finite. Thus, for large τ_0, and $\tau > \tau_0$, $|H_\mu(\tau) - H_\mu(\tau_0)| < Ar_\mu(\lambda)(\tau - \tau_0)$, for some constant $A > 0$, and $\tau_0 < \lambda < \tau$. □

COROLLARY 4.13. *On a collision orbit U_0, H_0, and T_0 have finite limits.*

Proof. On a collision orbit, where $t \to 0$, $\mathbf{z}$ tends to a point $\mathbf{z}^*$, and hence $U_0(\mathbf{x}, \mathbf{z})$ has a finite limit. Now, $H_0 = h - H_\Omega$, and $H_\Omega = \sum_\mu H_\mu$ has a finite limit by Theorem 4.12. Moreover, $T_0 = H_0 - U_0$. □

Remarks 4.14. We mentioned in the Introduction that one of the byproducts of the approach that we are following is that it provides the natural geometric setting and meaning to several estimates on collision orbits that appeared in the literature in the past. The result is that now these estimates are simple and obvious corollaries of Proposition 3.4, Theorem 4.2, and Shub's lemma. However, when some of these estimates were proved for the first time McGehee's transformation did not exist and obtaining one of these estimates required a lot of effort. For some of these estimates see [P-S], [S1], and [S].

(1) In Corollary 4.6 we saw that on collision orbits $r \approx bt^{2/3}$.

(2) A singularity as $t \to 0^-$ is a collision singularity if and only if $U \approx bt^{-2/3}$. The proof of the sufficiency is very easy and we omit it. *The necessity*: $U(\mathbf{x}, \mathbf{z}) = U_0(\mathbf{x}, \mathbf{z}) + U_\Omega(\mathbf{x}) = U_0(\mathbf{x}, \mathbf{z}) + r^{-1}V_\Omega(\mathbf{s})$. But $U_0(\mathbf{x}, \mathbf{z})$ is bounded on collision orbits (actually has a finite limit). Theorem 4.2 tells us that a collision orbit approaches a compact set of rest points. Hence, $\nabla V_\Omega(\mathbf{s}) \to 0$, and $V_\Omega(\mathbf{s})$ is bounded by Shub's lemma.

(3) If two masses m_i and m_j collide as $t \to 0$, $|q_i - q_j| = |x_i - x_j| = r|s_i - s_j| = t^{2/3}|s_i - s_j|$. Again, $|s_i - s_j|$, $i \neq j$, are bounded below by Shub's lemma.

4.15. Discussion. In § 3 we showed how to blow up the singular set $\Delta^*(\Omega)$, how to replace it by the collision manifold, and how to extend the vector field to that manifold. We also showed that, although the original vector field does not have rest points, the extended one has a set of rest points which lie on the collision manifold and which might not be finite; yet it is the union of a family of compact sets parametrized by the centres of mass positions and velocities $(\mathbf{z}, \dot{\mathbf{z}})$. In Corollary 4.13 we showed that $(\mathbf{z}, \dot{\mathbf{z}})$ has a finite limit on collision orbits. In this section (Theorem 4.12), we have shown that collision orbits approach the set of rest points. At this point we would like to mention some of the implications of this conclusion, and its more detailed form of Proposition 4.10, on the behaviour of the different clusters when they collapse simultaneously.

(1) The vector $\alpha = (\alpha_\mu \mid \mu \in \Omega)$ has norm one and its components represent the relative sizes of the different clusters. The fact that each α_μ has a limit that is bounded away from zero means that no cluster collapses infinitely faster than the others. More precisely, the ratio between the rates of collapse of any two of them is finite and not zero. Indeed,

$$\frac{r'_\mu}{r'_\nu} = \frac{u_\mu}{u_\nu}\left[\frac{\alpha_\nu}{\alpha_\mu}\right]^{1/2} \to \frac{\alpha^*_\mu}{\alpha^*_\nu}.$$

(2) For a cluster μ, the only variable that might not have a limit is $\mathbf{s}_\mu$. Yet, it converges to the set of critical points of $V_\mu = U_\mu | \Sigma^\mu$, $\Sigma^\mu = (\mathbf{S}^{3|\mu|-4} \setminus \Delta)$. But this set is compact and bounded away from $\Delta_\mu = \{\mathbf{s}_\mu | s_i = s_j \text{ for some } i \neq j \text{ in } \mu\}$. Thus, as the cluster μ collapses, no subcluster of μ collapses in the limit.

(3) It follows from the above that the set of rest points, and hence the ω-limit set of any collision orbit, stays away from the closure of the singular sets $\Delta^*(\Omega'')$, for all partitions Ω'' that are obtained by further partitioning Ω. This means that the stable set of $\Delta^*(\Omega)$ lies inside $\mathbf{E}_0$, which means that the choice of $\mathbf{E}_0$ was suitable for studying the asymptotic behaviour of collision orbits. Similar conclusions can be obtained for ejection orbits by letting $t \to -t$.

(4) In order to understand the asymptotic behaviour of the configuration $\mathbf{s}_\mu$ of a cluster μ, we need to understand the set of critical points of V_μ. This set need not be finite. In fact, since V_μ is invariant under the diagonal SO (3) action on Σ^μ, it follows that if $\mathbf{s}_\mu$ is a critical point of V_μ, its orbit under this action consists of critical points. We have denoted such an orbit by Orb $(\mathbf{s}_\mu)$ and called it a *central configuration.* By Shub's lemma, all these orbits lie in a compact subset of Σ^μ. Since SO (3) is compact, each of these orbits is a compact submanifold of Σ^μ. Nevertheless, there might be infinitely many of them.

(5) We mentioned, after defining it, that if the ω-limit set of a connected set lies in a compact set, it is also connected. Thus, if a certain central configuration is isolated, either $\mathbf{s}_\mu$ accumulates on it or it stays away from it in the limit, that is, for large τ. If all of them are isolated (or equivalently, if they are finitely many, since they lie in a compact subset), $\mathbf{s}_\mu$ approaches one and only one of them.

(6) We will devote the next three sections to showing that, whether or not $\mathbf{s}_\mu$ accumulates on only one of them, it does not go on rotating, that is, none of the clusters enters in an infinite spin as it collapses. This is known as the *Painlevé-Wintner problem* [S-H], [S], [W]. In fact, Saari and Hulkower showed in [S-H] that this cannot happen in the case of a total collapse, that is, when all of the n particles form one cluster and collide. The angular momentum in that case is identically zero, while for a cluster μ it goes to zero but does not vanish identically. We will follow their approach to prove the assertion for all cases except one case in which the cluster might approach a collinear central configuration. But collinear central configurations are nice since they are isolated as Conley indicated [Pc]. In order to cover this case we will study the subset of collinear configuration in Σ^μ, and apply the centre manifold theorem to the collinear critical points. Studying the manifold of collinear configurations is of interest by itself, since it gives more geometrical insight into the problems, which enables us to unify the classical theorem of Moulton about the number of collinear central configurations and Conley's result mentioned above, and stated in Lemma 6.11 below.

(7) In the statement of Theorem 4.1 of [S, p. 312], the author answers the Painlevé-Wintner question negatively for all cases. The proof that is given does not cover certain cases, namely, when one cluster ν approaches a collinear central configuration while another cluster μ accumulates on a component of $A_\mu = \{\mathbf{s} | \nabla V_\mu(\mathbf{s}) = 0, \mathbf{s} \text{ is not collinear}\}/\text{SO (3)}$, which is not a submanifold, or if the Hessian of V_μ is degenerate at A_μ. The author defers the proof of the case when one of the clusters tends to a collinear central configuration to Theorem 4.2 which deals with what he called "sufficiently hyperbolic sets" [S, Def. 3.1, p. 307]. However, the discussion supporting Theorem 4.2 depends on the condition that three certain related sets are submanifolds. These three sets are: $CC = \prod_{a=1}^{k} CC_a$ [S, pp. 304-305], "the union of components of CC which are sufficiently hyperbolic compact submanifold" denoted by *sh* [S, Def. 3.1, p. 307],

and a component $\mathcal{V}$ of *sh* that "corresponds to the choice of the limiting configurations of the various collisions" [S, p. 317, first paragraph]. The author treats them as manifolds and talks about the tangent space of some of their components. On page 304 the author mentions correctly that each CC_a is only real analytic, which is not necessarily a manifold. Now, both *sh* and $\mathcal{V}$ are subsets of CC, and hence are not necessarily submanifolds of the original space either, in particular, when the set A_μ is not a submanifold of Σ^μ or when the Hessian of V_μ is degenerate, i.e., in the cases we mentioned above. For $\mathcal{V}$ to satisfy § 3.1 it must be a manifold, and for this to happen, not only the component of $\mathcal{V}$ corresponding to the collinear configuration but also all the other components must be manifolds. Thus, the proof does not apply to this problematic case. We will prove the general case directly and without having to impose any conditions such as sufficient hyperbolicity, which are satisfied by "all currently known types of possible collision behavior." Those types are easier to handle on a case-by-case basis.

5. The action of the group SO (3) and the problem of infinite spin. Before presenting the ideas that we mentioned at the end of § 4, we need to collect some standard material about the action of Lie groups on manifolds. We will limit ourselves to the minimum that we need without going into generalities. Our main reference is [A-M].

5.1. Definitions, notation, and facts.

(1) Let G be a Lie group and M a manifold. Let $A: G\times M\to M$ be a C^∞ action of G on M. If $A(g,x)$ is denoted by $(g\cdot x)$, we have

$$g_1\cdot(g_2\cdot x)=(g_1g_2)\cdot x,\quad e\cdot x=x,\quad g,g_1,g_2\in G,\quad x\in M,$$

where e is the identity element of G. Moreover, we assume that G acts *faithfully* (or *effectively*) on M; that is, if $g\cdot x=x$ for all x in M, then $g=e$.

(2) Let $G\cdot x=\{g\cdot x\,|\,g\in G\}$ be the orbit of $x\in M$ under the action of G. If G is compact, $G\cdot x$ is a closed submanifold of M and its tangent space at x is given by

$$T_x(G\cdot x)=\{\xi_M(x)\,|\,\xi\in\mathfrak{g}\},\qquad \xi_M(x)=\frac{d}{dt}[(\exp t\xi)\cdot x]_{t=0},\tag{5.1.1}$$

where $\mathfrak{g}$ is the Lie algebra of G, and ξ_M is the vector field that ξ induces on M.

(3) Assume M is a Riemannian manifold with a metric denoted by $\langle\cdot,\cdot\rangle$. Then the tangent space of M at a point $x\in M$ is the orthogonal sum, $T_xM=P_x\oplus N_x$, where

$$P_x=T_x(G\cdot x)\quad\text{and}\quad N_x=\{Y\in T_xM\,|\,\langle Y,X\rangle=0\text{ for all }X\text{ in }P_x\}.\tag{5.1.2}$$

The dimension of P_x, and hence that of N_x, do not have to be equal for all x, for $G\cdot x$ is diffeomorphic to the quotient G/G_x, where $G_x=\{g\in G\,|\,g\cdot x=e\}$ is the stabilizer of x.

5.2. SO (3) action on $\mathbb{R}^{3n}$.

(1) Let $G=\mathrm{SO}(3)=\{\Theta\,|\,\Theta\ 3\times3\text{ matrix},\ \Theta^\dagger\Theta=I,\ \det\Theta=1\}$. Then, its Lie algebra $\mathfrak{g}$ is the set of all skew symmetric 3×3 matrices, $\Xi^\dagger=-\Xi$. The Lie algebra $\mathfrak{g}$ can be, and will be, identified with $\mathbb{R}^3$ via

$$\Xi=\begin{bmatrix}0&-x_3&x_2\\x_3&0&-x_1\\-x_2&x_1&0\end{bmatrix}\leftrightarrow\xi=\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}\in\mathbb{R}^3,$$

$$[\Xi_1,\Xi_2]\leftrightarrow\xi_1\times\xi_2,\qquad \Xi z=\xi\times z,\quad z\in\mathbb{R}^3.\tag{5.2.1}$$

From now on we shall assume that $G=\mathrm{SO}(3)$ and $\mathfrak{g}\cong\mathbb{R}^3$.

(2) The group SO (3) acts effectively on $\mathbb{R}^{3n}$ by

$$g(x_1, x_2, \cdots, x_n) = (gx_1, gx_2, \cdots, gx_n), \qquad g \in G, \quad x_i \in \mathbb{R}^3, \quad i = 1, 2, 3, \cdots, n.$$

This action leaves $\mathbf{S}^{3n-1}$ invariant. Furthermore, it leaves the following two sets invariant,

$$S^{3n-4} = \left\{ s \in \mathbf{S}^{3n-1} \,\middle|\, \sum_i m_i s_i = 0 \right\} \quad \text{and} \quad \Sigma = \Sigma^{3n-4} = \mathbf{S}^{3n-4} \backslash \Delta.$$

Note that all the unit spheres mentioned here are defined in terms of the metric given by the matrix M, yet, $\|gs\|^2 = \sum_i m_i |gs_i|^2 = \|s\|^2$, since $|gs_i| = |s_i|$ for all i.

(3) Using (5.1.1) and the identification (5.2.1) we can decompose each $T_s\Sigma$ as in (5.1.2) as follows:

$$\mathbf{X}_s = T_s(G \cdot s) = \{\xi \wedge s \mid \xi \in \mathbb{R}^3 = \mathfrak{g}\},$$

where

$$\xi \wedge s = (\xi \times s_1, \xi \times s_2, \cdots, \xi \times s_n),$$

and

$$N_s = \left\{ Y \in \mathbb{R}^{3n} \,\middle|\, \sum_i m_i Y_i = 0, \langle Y, \eta \wedge s \rangle = 0 \text{ for all } \eta \in \mathbb{R}^3 \right\}.$$

5.3. The configuration of a cluster μ.

(1) Consider a fixed but arbitrary cluster μ. To simplify the notation in the next few articles we will drop the subscript μ and assume the cluster under consideration has n particles. Thus, for this cluster we have $(s, w) \in T_s\Sigma$, $\alpha \in (0, 1)$ has a limit that is bounded away from zero and one, and $w \to 0$ as $\tau \to \infty$, but not necessarily fast enough to ensure the convergence of s. We have, with all the variables below depending on τ,

$$s' = \alpha^{-3/2} w, \quad w = X + Y \quad \text{for some } X \in \mathbf{X}_s, \text{ and } Y \in N_s, \tag{5.3.1}$$

that is,

$$X = \xi \wedge s, \quad \langle Y, \eta \wedge s \rangle = 0 \quad \text{for some } \xi \in \mathbb{R}^3, \quad \text{for all } \eta \in \mathbb{R}^3, \tag{5.3.2}$$
$$\sum_i m_i Y_i = 0 \quad \text{and} \quad X, Y \to 0 \quad \text{as } \tau \to \infty.$$

Finally, we write each s_i as the sum of two components, one in the direction of ξ and the other in the normal direction, as in Fig. 5.1.

$$s_i = \beta_i \xi + z_i \quad \text{and} \quad (z_i, \xi) = 0, \quad i = 1, 2, 3, \cdots, n. \tag{5.3.3}$$

THEOREM 5.4. (i) *As* $\tau \to \infty$, $|\xi| \, \|z\|^2 \to 0$ *exponentially.*

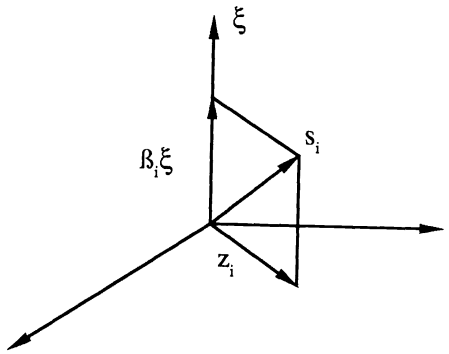


FIG. 5.1

(*ii*) *If ξ is bounded or if $\|z\|$ is bounded away from zero, $\|X\| = |\xi|\,\|z\| \to 0$ exponentially and s does not undergo an infinite spin as $\tau \to \infty$.*

Proof. (i) Let $\mathbf{c} = \mathbf{c}_\mu$ be the angular momentum of the cluster under consideration. Then,

$$\mathbf{a} = (\alpha r)^{-1/2}\mathbf{c} = \sum_i m_i(s_i \times w_i) = \sum_i m_i(s_i \times X_i) + \sum_i m_i(s_i \times Y_i). \tag{5.4.1}$$

The last sum is identically zero, for if η is any vector in $\mathbb{R}^3$, since $Y \in N_s$, we have

$$\left(\eta, \sum_i m_i s_i \times Y_i\right) = \sum_i m_i(\eta, s_i \times Y_i) = \langle \eta \wedge s, Y\rangle = 0. \tag{5.4.2}$$

Using (5.3.2) above, we obtain

$$\mathbf{a} = \sum_i m_i s_i \times X_i = \sum_i m_i s_i \times (\xi \times s_i) = -\|\xi\|^2\left(\sum_i m_i \beta_i z_i\right) + \|z\|^2 \xi. \tag{5.4.3}$$

Since ξ is orthogonal to each z_i, and since $\mathbf{c}$ satisfies the inequality (4.10.0), the norm of the last term satisfies the following inequality, which proves part (i):

$$|\xi|\,\|z\|^2 \leqq \|\mathbf{a}\| \leqq br^3 < b_1 \exp[3v_-\tau]. \tag{5.4.4}$$

(ii) On the other hand, since each z_i is orthogonal to ξ, the norm of X satisfies

$$\|X\|^2 = \|\xi \wedge s\|^2 = \|\xi \wedge z\|^2 = |\xi|^2\|z\|^2. \tag{5.4.5}$$

Thus, if either ξ is bounded or $\|z\|$ is bounded away from zero for large τ, $\|X\| \to 0$ exponentially fast and the cluster under investigation does not undergo an infinite spin as it collapses. □

COROLLARY 5.5. *An arbitrary cluster μ does not undergo an infinite spin as it collapses in the following three cases*:

(i) *If it does not approach a collinear central configuration,*

(ii) *If its particles lie in a fixed plane, i.e., in the planar problem, or*

(iii) *If it contains exactly two particles.*

Proof. (i) If the cluster μ stays away from collinear central configurations, then, for large τ, $\|z\|$ is bounded away from zero.

(ii) If the particles of the cluster are always in a fixed plane, they will always rotate around an axis that is perpendicular to that plane. In this case, $s_i = z_i$ and $\beta_i = 0$ for all i, which means that $\|z\| = \|s\| = 1 > 0$.

(iii) Let μ have exactly two particles. Since the origin of the cluster is at its centre of mass, we can assume s has only one vector component. Thus,

$$\mathbf{a} = -m|\xi|^2\beta z + \|z\|^2\xi = |\xi|\,\|z\|\left[m|\xi|\beta\frac{z}{\|z\|} + \frac{\|z\|\xi}{|\xi|}\right]. \tag{5.5.1}$$

Both $z/\|z\|$ and $\xi/\|\xi\|$ are unit vectors and $1 = |s|^2 = (\beta\xi)^2 + \|z\|^2$. Hence, if $z \to 0$, $\beta|\xi| \to 1$, and the norm of the square bracket in (5.5.1) goes to one. Thus, $|\xi|\,\|z\| \to 0$ exponentially fast. □

Remark 5.6. The only case that is not covered by Theorem 5.4 is when s is not bounded away from collinear central configurations. This is a consequence of the fact that the SO (3)-orbit of a collinear configuration is singular relative to the action of SO (3). More precisely,

$$\text{orb}\,(s) \cong \frac{\text{SO}\,(3)}{G_s}, \qquad G_s = \{g \in \text{SO}\,(3) \mid g \cdot s = s\}.$$

If s is not collinear, G_s consists of the identity element, but, if s is a collinear configuration along the line given by the unit vector $\underline{a} \in \mathbf{S}^2$, G_s consists of all the rotations around the axis given by $\underline{a}$, and hence diffeomorphic to S^1.

On the other hand, collinear central configurations are isolated according to Conley's Lemma 6.1 [CC], which shows that either $\mathbf{s}$ approaches a specific collinear central configuration or stays away from them. In the latter case, Corollary 5.5 shows that the cluster μ does not undergo an infinite spin. In order to resolve this problem in the former case, we shall study the set of all collinear configurations as a submanifold of Σ.

6. The manifold of collinear configurations. In this section we intend to study the set of collinear configurations of n particles in space. We shall show that it is a submanifold $\mathbf{L}$ of Σ, diffeomorphic to $(\mathbf{S}^2 \times \mathbf{S}^{n-2})/\sim$, where $(e, \underline{a}) \sim (-e, -\underline{a})$. We shall also write $T_s\Sigma$ as the orthogonal sum $T_s\Sigma = T_s\mathbf{L} \oplus Z_s$, $T_s\mathbf{L} = T_s(G \cdot s) \oplus Y_s$, for $s \in \mathbf{L}$, and $G = \mathrm{SO}(3)$. Then, we shall study the Hessian of V, where V is the restriction of the potential function to the unit sphere, and show that at an arbitrary collinear critical point (central configuration) $\mathbf{s}$ it is identically zero on $T_s(G \cdot \mathbf{s})$, negative definite on Y_s, and positive definite on Z_s. The first assertion is expected since V is constant on the SO(3) orbits. The second will be shown after writing the Hessian explicitly, and will give, as a corollary, Moulton's theorem (which states that there are $n!/2$ collinear central configurations). The idea of the proof of the third is due to Conley [Pc]. From that analysis of the Hessian of V at such a critical point, it follows that the collinear ones are isolated. It also means that the SO(3) orbit of a collinear critical point is normal hyperbolic which will enable us to apply the centre manifold theorem and show that if $\mathbf{s}$ approaches a collinear central configuration, it must have a limit, which means that it does not undergo an infinite spin and settles this question.

6.1. Notation and definitions. Let Σ be Σ^{3n-4} as defined in § 5.2, and let $\underline{m} = (m_1, m_2, \cdots, m_n)$. Define the following:

$$\Lambda = \{\underline{a} \in S^{n-1} \mid (\underline{a}, \underline{m}) = 0,\ a_i \neq a_j \text{ when } i \neq j\} \approx \mathbf{S}^{n-2} \backslash \Delta, \tag{6.1.1}$$

$$\begin{aligned} \mathbf{L} &= \{s \in \Sigma \mid s \text{ is collinear}\} \\ &= \{s \in \Sigma \mid \text{there are } e \in \mathbf{S}^2 \text{ and } \underline{a} \in \Lambda, \text{ such that } s_i = a_i e \text{ for all } i\}, \end{aligned} \tag{6.1.2}$$

$$\mathbb{L} = \frac{\mathbf{S}^2 \times \Lambda}{\sim} \quad \text{where } (e, \underline{a}) \sim (f, \underline{b}) \Leftrightarrow (f, \underline{b}) = \pm(e, \underline{a}), \tag{6.1.3}$$

$[e, \underline{a}]$ = the equivalent class of $(e, \underline{a})$.

PROPOSITION 6.2. *L is diffeomorphic to $\mathbb{L}$.*

Proof. Define $\phi : \mathbb{L} \to \Sigma$ by $\phi([e, \underline{a}]) = \mathbf{s}$, $s_i = a_i e$, for $i = 1, 2, \cdots, n$. Then, ϕ is well defined, injective, and $\phi(\mathbb{L}) = \mathbf{L}$. Actually, ϕ defines a manifold structure on $\mathbf{L}$ by carrying that of $\mathbb{L}$ to it. Thus, what we need to show is that ϕ is an embedding. In order to do that we compute $\phi_* : T\mathbb{L} \to T\Sigma$. Near a point $[e, \underline{a}] \in \mathbb{L}$, $(e, \underline{a})$ form a coordinate system. Let $(X, Y)_{\pm} = \{(X, Y), -(X, Y)\}$ be in $T_{[e,a]}\mathbb{L}$. Then,

$$(X, e) = 0, \quad \langle Y, \underline{a} \rangle = 0, \quad (Y, \underline{m}) = 0,$$

and

$$\phi_*([e, \underline{a}]) \begin{bmatrix} X \\ Y \end{bmatrix} = \frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} \phi((e, a) + \varepsilon(X, Y)) = \begin{bmatrix} a_1 X + Y_1 e \\ a_2 X + Y_2 e \\ \vdots \\ a_n X + Y_n e \end{bmatrix}.$$

This is well defined, since if we choose $(-e, -\underline{a})$ as a coordinate system near $[e, \underline{a}]$, we replace (X, Y) by $(-X, -Y)$. Let (X, Y) be in the kernel of $\phi_*([e, \underline{a}])$. Since X and e are orthogonal, it follows that $a_i X = 0$, $Y_i e = 0$, for $i = 1, 2, \cdots, n$. But $\|\underline{a}\| = |e| = 1$. Thus, $X = 0$ and $Y = 0$, and hence $\phi_*([e, \underline{a}])$ is an isomorphism. Since the domain and range of ϕ are compact, ϕ is a homeomorphism in the topology that $\mathbf{L}$ inherits from Σ. Hence, it is an embedding. □

PROPOSITION 6.3. *$\mathbb{L}$, and hence $\mathbf{L}$, have $(n!/2)$ simply connected components.*

Proof. $\mathbf{S}^2 \times \Lambda$ has $n!$ components, one for each different ordering, $a_{\sigma(1)} < \cdots < a_{\sigma(n)}$, where σ is a permutation of n objects. If we take the quotient, we get $(n!/2)$ components. □

PROPOSITION 6.4. *Let $\mathbf{s} = \phi([e, \underline{a}]) \in \mathbf{L}$. Then, $G \cdot s \subset \mathbf{L}$. Moreover, if we apply $\phi_*([e, \underline{a}])$ to $X = (\xi \times e) \in T_e \mathbf{S}^2$ and $\underline{b} \in T_{\underline{a}} \Lambda$, we find that $T_s \mathbf{L} = T_s(G \cdot \mathbf{s}) \oplus \mathbf{Y}_s$, where*

$$T_s(G \cdot s) = \{X \mid X_i = a_i(\xi \times e) \text{ for some } \xi \in \mathbb{R}^3\},$$

$$\mathbf{Y}_s = \{Y \mid Y_i = b_i e \text{ for some } b \text{ with } \langle \mathbf{s}, Y \rangle = \langle \underline{a}, \underline{b} \rangle = 0 \text{ and } (\underline{m}, \underline{b}) = 0\}.$$

It follows that $\langle X, Y \rangle = (X, Y) = 0$, as $(e, \xi \times e) = 0$. Moreover, $\dim \mathbf{X}_s$ is constant for $\mathbf{s} \in \mathbf{L}$. □

We shall identify X with $(X, 0)^\dagger$, and Y with $(0, Y)^\dagger$.

DEFINITION 6.5. Let $(T\Sigma)|\mathbf{L} = \mathbf{X} \oplus \mathbf{Y} \oplus \mathbf{Z}$, where $T\mathbf{L} = \mathbf{X} \oplus \mathbf{Y}$, and $\mathbf{X}_s = T_s(G \cdot \mathbf{s})$, for $\mathbf{s} \in \mathbf{L}$. The direct sums are orthogonal relative to the inner product given by the matrix M.

COROLLARY 6.6. *If $\mathbf{s} = \phi([e, \underline{a}]) \in \mathbf{L}$ and $Z \in \mathbf{Z}_s$ then, for all $\xi \in \mathbb{R}^3$ and for all $Y \in \mathbf{Y}$,*

$$\left\langle \sum_i m_i a_i Z_i, \xi \times e \right\rangle = \langle Z, X \rangle = 0, \tag{6.6.1}$$

$$\left\langle \sum_i m_i b_i Z_i, e \right\rangle = \langle Z, Y \rangle = 0, \tag{6.6.2}$$

$$(e, Z_i - Z_j) = 0, \quad \text{for all } i, j. \tag{6.6.3}$$

Proof. The first two identities follow immediately from writing X and Y explicitly. We prove the third for $i = 1$ and $j = 2$. Let $W = (m_2 e, -m_1 e, 0, \cdots, 0)^\dagger$, and $Y = W - f\mathbf{s}$, where $f = \langle W, \mathbf{s} \rangle$. Then, $\sum_i m_i Y_i = 0$, and $\langle Y, \mathbf{s} \rangle = 0$, and hence, $Y \in \mathbf{Y}$. Since $\langle Z, \mathbf{s} \rangle = 0$ for $Z \in T_s \Sigma$, it follows that $0 = \langle Y, Z \rangle = \langle Y + f\mathbf{s}, \mathbf{Z} \rangle = \langle \mathrm{W}, \mathrm{Z} \rangle = \mathrm{m}_1 \mathrm{m}_2 (\mathrm{e}, \mathrm{Z}_1 - \mathrm{Z}_2)$. □

PROPOSITION 6.7. *Let h_s be the quadratic form associated with the Hessian of the potential $V : \Sigma \to \mathbf{R}$, at a critical point $\mathbf{s} \in \Sigma$, and let $w \in T_s \Sigma$. Then,*

$$h_s(w) = HV(\mathbf{s})(w, w) = \langle w, Bw \rangle + k_s(w) + V(\mathbf{s}) \|w\|^2, \tag{6.7.1}$$

$$k_s(w) = -3 \sum_{i<j} \frac{m_i m_j}{|s_i - s_j|^5} (s_i - s_j, w_i - w_j)^2, \tag{6.7.2}$$

where

$$B = M^{-1} A, \qquad \langle w, Bw \rangle = (w, Aw) = \sum_{i<j} \frac{m_i m_j}{|s_i - s_j|^3} |w_i - w_j|^2 \geqq 0, \tag{6.7.3}$$

with

$$A_{ij} = \begin{cases} -\dfrac{m_i m_j}{|s_i - s_j|^3} & \text{for } i \neq j, \\ \displaystyle\sum_{k \neq i} \frac{m_i m_k}{|s_i - s_k|^3} & \text{for } i = j, \end{cases} \tag{6.7.4}$$

$$(f, Af) = \sum_{i<j} \frac{m_i m_j}{|s_i - s_j|^3} (f_i - f_j)^2 \quad \text{for any vector } f \in \mathbb{R}^n. \tag{6.7.5}$$

□

COROLLARY 6.8. (1) $\mathbf{s} \in \Sigma$ *is a critical point of* V *if and only if* $A(\mathbf{s})\mathbf{s} = DV(\mathbf{s})$, *i.e.*, $B(\mathbf{s})\mathbf{s} = -V(\mathbf{s})\mathbf{s}$.

(2) *If* $s \in \Sigma$ *is a critical point of* V, *and* $X \in T_s(G \cdot \mathbf{s})$, *then* $k_s(X) = 0$ *and* $h_s(X) = 0$.

Proof. (1) This assertion follows from (6.7.5) and the fact that $\mathbf{s}$ is a critical point of V if and only if $M^{-1}DV(\mathbf{s}) = -V(\mathbf{s})\mathbf{s}$.

(2) The components of X are of the form $X_i = \xi \times s_i$, $\xi \in \mathbb{R}^3$. It follows that for all i and j $(s_i - s_j, X_i - X_j) = (s_i - s_j, \xi \times (s_i - s_j)) = 0$. Thus, $k_s(X) = 0$. Direct calculations show that $AX = A(\xi \wedge \mathbf{s}) = \xi \wedge A\mathbf{s}$, and $MX = M(\xi \wedge \mathbf{s}) = \xi \wedge M\mathbf{s}$. Using (1) we get the following, which implies that $h_s(X) = 0$:

$$(X, AX) = (\xi \wedge \mathbf{s}, \xi \wedge A\mathbf{s}) = -V(\mathbf{s})(\xi \wedge s, \xi \wedge M\mathbf{s}),$$

$$V(\mathbf{s})\|X\|^2 = V(\mathbf{s})(\xi \wedge \mathbf{s}, \xi \wedge M\mathbf{s}) = -(X, AX) = -\langle X, BX\rangle. \qquad \square$$

Remark 6.9. For the rest of this section we will consider collinear critical points. Thus, we shall only pay attention to one component of $\mathbf{L}$, say Γ_0. We can assume it has a coordinate system $(e, \underline{a})$, such that $\underline{a} \in \Gamma$,

$$(6.9.1) \qquad \Gamma = \{\underline{x} \in \mathbf{S}^{n-1} \mid \langle \underline{x}, \vec{\mathbf{1}}\rangle = 0, x_1 < x_2 < \cdots < x_n\}, \qquad \vec{\mathbf{1}} = (1, 1, \cdots, 1) \in \mathbb{R}^n.$$

Note that both the vector $\vec{\mathbf{1}}$ and the inner product $\langle \cdot, \cdot \rangle$ do not depend on the ordering of the a_i's, and hence, we do not lose any generality by our choice of Γ. Moreover, Γ_0 can be identified with $\mathbf{S}^2 \times \Gamma$, and we will write $\phi(e, \underline{a})$ instead of $\phi([e, \underline{a}])$.

THEOREM 6.10. *Let* $\mathbf{s} = \phi([e, \underline{a}])$ *be a collinear critical point for* V *and* $w \in T_s\Sigma$. *Then,*

$$(6.10.1) \qquad h_s(w) = \sum_{i<j} \frac{m_i m_j}{|a_i - a_j|^3} |w_i - w_j|^2 - 3 \sum_{i<j} \frac{m_i m_j}{|a_i - a_j|} (e, w_i - w_j)^2 + V(\mathbf{s})\|w\|^2,$$

(6.10.2) $h_s | \mathbf{Y}$ *is negative definite, that is,* $h_s(Y) < 0$, *for all* $Y \in \mathbf{Y}$, *and*

(6.10.3) $k_s(Z) = 0$ *for all* $Z \in \mathbf{Z}_s$.

Proof. The first identity follows from (6.7.1) by writing $\mathbf{s}$ in terms of e and $\underline{a}$. To prove the second we take $Y \in \mathbf{Y}$ with $Y_i = b_i e$ for some $\underline{b}$ satisfying $\langle \underline{a}, \underline{b}\rangle = 0$, $(\underline{m}, \underline{b}) = 0$. Direct calculations show that $k_s(Y) = -3(Y, AY)$. Thus, $h_s(Y) = -2(Y, AY) + V(\mathbf{s})\|Y\|^2 < 0$. The last identity follows from (6.6.3) and from observing that k_s is nothing but the second sum in the first identity. $\square$

6.11. Conley's lemma for collinear central configurations. *Let* $\mathbf{s} = \phi([e, \underline{a}])$ *be a collinear critical point for* V. *Then, for all* $Z \in \mathbf{Z}_s$, $h_s(Z) > 0$.

Remark 6.2. What this assertion means is that at a collinear central configuration, h_s is positive definite in the direction normal to the submanifold $\mathbf{L}$ of collinear configurations. The idea of the proof is due to Conley [Pc]. We shall present the proof of this theorem in a series of Lemmas A-D.

LEMMA 6.13. A. *Let* $Z \in \mathbf{Z}_s$; *then by* (6.10.3), $k_s(Z) = 0$. *Thus,*

$$h_s(Z) = \langle Z, BZ\rangle - \lambda\|Z\|^2, \qquad \lambda = -V(\mathbf{s}). \qquad \square$$

Observation 6.14. To prove Conley's lemma it is enough to show that all the nonzero eigenvalues of B are strictly bigger than λ. In the next few sections we shall study B as an $n \times n$ matrix acting of $\mathbb{R}^n$.

LEMMA 6.15. B. (1) *B has zero as an eigenvalue whose eigenspace is* $[\vec{\mathbf{1}}] = \text{span}\{\vec{\mathbf{1}}\}$.

(2) *If* $\mu \neq 0$ *is an eigenvalue of* B, *then* $\mu > 0$.

(3) *The subspace normal to* $\vec{\mathbf{1}}$ *is given by* $\mathbf{P}=\{x\in\mathbb{R}^n \mid \sum_i m_i x_i = 0\}$.

(4) *The vector* $\underline{a}$ *is an eigenvector for* B *with eigenvalue* λ. *Moreover, it lies in* $\mathbf{P}$.

Proof. (1) Recall that

$$\langle Z, BZ\rangle = (Z, AZ) = \sum_{i<j} \frac{m_i m_j}{|a_i - a_j|^3}|Z_i - Z_j|^2.$$

This expression vanishes if and only if $Z_1 = Z_2 = \cdots = Z_n$, i.e., Z belongs to $[\vec{\mathbf{1}}]$.

(2) If $\mu \neq 0$ and $BZ = \mu Z$, then $Z \notin [\vec{\mathbf{1}}]$, and $\mu\|Z\|^2 = \langle Z, BZ\rangle > 0$.

(3) Since $\langle x, \vec{\mathbf{1}}\rangle = \sum_i m_i x_i$, the assertion follows.

(4) From Corollary 6.8(2), and the fact that each $s_i = a_i e$, it follows that $B\underline{a} = \lambda \underline{a}$. The vector $\underline{a}$ is in $\mathbf{P}$ because $0 = \sum_i m_i s_i = (\sum_i m_i a_i)e$, and $|e| = 1$. □

COROLLARY 6.16. *Let* β *be the bilinear form given by* $\beta(x) = \langle x, Bx\rangle$ *for* $x \in \mathbb{R}^n$. *Let* $\alpha > 0$, *and define*

$$C_\alpha = \{x \mid \beta(x) = \alpha\}, \qquad E_\alpha = C_\alpha \cap P.$$

Then, the following follows from the previous lemma:

(1) *The surface* C_α *is a cylinder with axis along the vector* $\vec{\mathbf{1}}$.

(2) *Each* E_α *is an ellipsoid.*

(3) *Furthermore, since* $\underline{a}$ *is an eigenvector of* B *with eigenvalue* λ, *and since it belongs to* $\mathbf{P}$, *it is one of the vertices of* E_λ.

(4) $\underline{a} \in E_\lambda \cap \Gamma$, *and* E_λ *is tangent to* Γ *at* $\underline{a}$. □

COROLLARY 6.17. *Let* μ *be an eigenvalue of* B, *different from zero and* λ, *with a unit eigenvector* $\underline{b}$. *Then,* $\underline{b}$ *does not belong to an open neighbourhood of the closure of* Γ.

Proof. We know that $\underline{a} \in \Gamma$. Since $\lambda \neq \mu$ and B is symmetric with respect to the inner product $\langle \cdot, \cdot \rangle$, it follows that $\langle \underline{a}, \underline{b}\rangle = 0$. Moreover, $\underline{b}$ does not belong to $[\vec{\mathbf{1}}]$, and hence, its components are not all equal. Since $\langle \underline{a}, \vec{\mathbf{1}}\rangle = \langle \underline{b}, \vec{\mathbf{1}}\rangle = 0$, it follows that

$$0 = \kappa\langle \underline{a}, \underline{b}\rangle = \kappa \sum_i m_i a_i b_i = \sum_{i<j} m_i m_j (a_i - a_j)(b_i - b_j) \quad \text{with } \kappa = \sum_i m_i.$$

But each $(a_i - a_j)$ is negative since $\underline{a} \in \Gamma$. Hence, unless $(b_i - b_j) > 0$ for some $i < j$, the right-hand side does not vanish. Since there are only finitely many eigenvectors, $\underline{a}$ is the only one in a neighbourhood of the closure of Γ except perhaps other eigenvectors of λ if they exist. □

LEMMA 6.18. C. *Let* $\mathbf{S}_1 = \{x \in \mathbf{S}^{n-1} \mid \langle x, \vec{\mathbf{1}}\rangle = 0\} \approx \mathbf{S}^{n-2}$, *and* $g = \beta|\mathbf{S}_1$. *Consider the gradient system defined on* $\mathbf{S}_1$ *by*

$$\frac{dx}{dt} = \nabla g(x) = Dg(x) - \langle Dg(x), x\rangle x = 2[Bx - g(x)x]. \tag{6.18.1}$$

(a) $dg/dt \geqq 0$, *and vanishes only at rest points, that is,* g *is a Lyapunov function.*

(b) *The flow is transversal to the boundary of* Γ, *which we denote by* $\partial\Gamma$.

(c) *The flow has a unique fixed point at* $\underline{a}$, *which is also an extremal point of* g.

(d) *The fixed point* $\underline{a}$ *is a repeller and* g *has a global minimum at* $\underline{a}$ *and is increasing everywhere else.*

(e) $E_\lambda \cap \Gamma = \{\underline{a}\}$.

Proof. (a) $dg/dt = \|\nabla g(x)\|^2 \geqq 0$, and vanishes only at fixed points.

(b) Let $x \in \partial\Gamma$, that is, $x_k = x_{k+1}$, and $x_l \neq x_{l+1}$, for some integers $k \neq l$ between 1 and n. Moreover, $x_1 \leqq \cdots \leqq x_n$, and $a_1 < \cdots < a_n$. Thus, each term in the following

sum is nonnegative and the sum is positive:

$$\frac{d}{dt}(x_k - x_{k+1})|_{x_k = x_{k+1},} = \sum_{j \neq k,k+1} m_j(x_k - x_j)\left[\frac{1}{|a_j - a_k|^3} - \frac{1}{|a_j - a_{k+1}|^3}\right] > 0.$$

Thus, the flow is transversal to the boundary of Γ, and leaves it as time progresses.

(c) A point x in Γ is a rest point if and only if it is an eigenvector of B with eigenvalue $\mu \neq 0$, since $\ker B = [\vec{1}]$. If $\mu \neq \lambda$, we conclude from Corollary 6.17 that x could not belong to Γ. If $\mu = \lambda$, and $x \notin \operatorname{span}\{\underline{a}\}$, it follows that the eigenspace of λ is at least two-dimensional, and hence, it must intersect $\partial\Gamma$, and we can assume that x is in $\partial\Gamma$. But we have just seen that $\partial\Gamma$ contains no fixed points.

(d) The flow leaves $\partial\Gamma$ as t increases. Thus, for any $p \in \operatorname{cl}(\Gamma)$, the α-limit of p, denoted by $\alpha(p)$, is an invariant compact subset of the interior of Γ. Since the flow is gradient, and since $\underline{a}$ is the only fixed point there, $\alpha(p) = \{\underline{a}\}$. Hence, p is a repeller. Thus, g takes its minimum at $\underline{a}$.

(e) We have already seen in Corollary 6.16(4) that $\underline{a} \in E_\lambda \cap \Gamma$. Thus, g assumes its minimum on $E_\lambda \cap \Gamma$, and hence, $E_\lambda \cap \Gamma$ consists of rest points, but we have only one, $\underline{a}$. □

Finally we give the last lemma in the proof of Conley's Lemma 6.11.

LEMMA 6.19. D. *The nonzero eigenvalues of B are strictly larger than* λ.

Proof. Let E be an ellipsoid defined by a quadratic form B_0 with positive eigenvalues $0 < \lambda_1 \leqq \cdots \leqq \lambda_m$,. Let $v_1, \cdots, v_m$ be in E and be eigenvectors corresponding to $\lambda_1, \cdots, \lambda_m$, respectively. Then, $v_1, \cdots, v_m$ are vertices for E and $|v_1| \geqq \cdots \geqq |v_m|$. Now, we apply this observation to the restriction $B|P$ in order to show that λ is the smallest nonzero eigenvalue of B. What we need to show is that $\underline{a}$ is the longest axis. But we know that Γ is an open subset of the unit sphere S_1, and that it is tangent to E_λ at $\underline{a}$. Actually, we have seen in Lemma 6.18.C(d) that $E_\lambda \cap \Gamma = \{\underline{a}\}$. Thus, S_1 lies either inside E_λ or outside it, and hence, $\underline{a}$ is either the shortest or the longest axis, respectively. If it was the shortest, E_λ would be tangent to Γ from outside, and hence, $E_{\lambda+\varepsilon} \cap \Gamma$ would be empty for all $\varepsilon > 0$. But we know that g takes its minimum value λ at $\underline{a}$. Thus, this intersection cannot be empty for small ε, which implies that $\underline{a}$ is strictly the longest axis. □

6.20. Important remark. Conley's lemma, whose proof has just been completed, implies that collinear central configurations are isolated. Thus, as we mentioned in the discussion in § 4.15 above, a cluster μ either approaches one and only one collinear central configuration or accumulates on the noncollinear ones. It was shown in Corollary 5.5 that in the latter case, the cluster μ will not undergo an infinite spin as it collapses. Now, since we know that each collinear central configuration is isolated, we can study the asymptotic behaviour of μ in the former case and resolve the last part of the question of infinite spin.

COROLLARY 6.21 (Moulton's theorem). *There are exactly* $(n!/2)$ *collinear central configurations.*

Proof. Recall that a collinear central configuration is an SO (3)-orbit of a collinear critical point of V. Moreover, we saw in Proposition 6.3 that $\mathbf{L}$ has $(n!/2)$ simply connected components. Thus, it is enough to consider the component $\Gamma_0 \approx S^2 \times \Gamma$, and consider V as a function of x, where x belongs to Γ which is diffeomorphic to D^{n-2}, the open ball of dimension $n-2$. By (6.10.2), each critical point is nondegenerate and a local maximum; hence, an isolated repeller and its Morse index is $(n-2)$. Shub's Lemma 3.6, implies that the critical points lie in a compact subset of D^{n-2}, hence, they are finitely many. Moreover, $V \to -\infty$ as x approaches the boundary of Γ. Consider

the gradient flow given on Γ by

$$\dot{x} = -\nabla V(x), \qquad x \in \Gamma \approx D^{n-2}. \tag{6.21.1}$$

For $c<0$, define the set $A^c = \{x \in \Gamma \mid V(x) \leqq c\}$, whose boundary is $B^c = V^{-1}(c)$. For sufficiently small $c<0$, all the critical points lie in the interior of A^c. Thus, B^c does not contain any critical points, and hence is a closed regular submanifold of A, and hence compact. It has codimension one. Moreover, V is strictly decreasing on solutions which are not rest points. Thus, the flow is transversal, actually, normal to B^c and points outward. Furthermore, the α-limit set of B^c is a subset of A^c, and hence, is nonempty, compact, connected, and invariant. As $t \to -\infty$, V increases and is bounded above. Hence, for a point x in B^c, $\alpha(x)$ belongs to a level curve of V, and since V is strictly decreasing on solutions which are not rest points, $\alpha(x)$ is actually a rest point. Now, $\alpha(B^c)$ is a subset of the set of rest points. But it is connected, and the set of rest points is finite. Thus, $\alpha(B^c)$ is actually a rest point, and hence, there is a unique fixed point in Γ. □

7. The problem of infinite spin for collinear configurations.

7.1. We begin this section by recalling the differential equation that governs the motion of a fixed cluster μ which approaches a collinear central configuration. We let $r_\mu = r\alpha_\mu$, and since we are concerned with a certain cluster, and since α_μ goes to a positive limit as $\tau \to \infty$, we rescale time by $\alpha_\mu^{3/2}$ and still denote it by τ and use a dash to denote differentiation with respect to it. We will also omit the subscript μ for simplicity:

(III)
$$r' = ru, \qquad \mathbf{s}' = \mathbf{w},$$

$$u' = \frac{u^2}{2} + \|\mathbf{w}\|^2 + V(\mathbf{s}) + r^2 f = \frac{\|\mathbf{w}\|^2}{2} + K + r^2 f,$$

$$\mathbf{w}' = -\frac{u}{2}\mathbf{w} - \|\mathbf{w}\|^2 \mathbf{s} - \nabla V(\mathbf{s}) + r^2 \mathbf{g},$$

$$K = \frac{1}{2}(u^2 + \|\mathbf{w}\|^2) + V(\mathbf{s}) = rH.$$

Remark 7.2. The two functions f and $\mathbf{g}$ are real analytic at $r=0$, and represent the effect of the rest of the particles on the cluster μ. We can consider them as C^∞ functions of the time τ and consider (III) as the differential equation governing the motion of n particles moving under their mutual gravitational force and subjected to an external force given by $r^2 f$ and $r^2 \mathbf{g}$. In this case the system (III) is no longer autonomous. In order to recover this property we define a new variable β and enlarge the system by adding β to it. We also recall that the energy function H is bounded, but not necessarily constant, and hence K vanishes on $r=0$.

COROLLARY 7.3. *Let $\beta : (-\infty, \infty) \to (0, +\infty)$ be given by $\beta(\tau) = \exp[-\nu\tau]$, for some positive ν to be determined soon. The mapping β is a diffeomorphism and hence we can, and will, consider f and $\mathbf{g}$ as functions of β. Moreover, $\beta \to 0$ as $\tau \to \infty$, and the two functions, $\tilde{f}(\beta) = r(\beta) f(\beta)$, and $\tilde{\mathbf{g}}(\beta) = r(\beta)\mathbf{g}(\beta)$ can be extended to a C^k at $\beta = 0$ by choosing ν small enough, with $\beta' = 0$. Hence β can be extended to C^k even functions through $\beta = 0$.*

Proof. From Corollary 4.6 there are $v_1, v_2 > 0$, and $b_1, b_2 > 0$, such that, for large τ, $b_1 \exp(-v_1\tau) < r(\tau) < b_2 \exp(-v_2\tau)$. Moreover,

$$r\frac{df}{d\beta} = -r\frac{df}{d\tau}\beta^{-1}, \qquad \frac{dr}{d\beta} = -\frac{ru}{v\beta},$$

and u has a nonzero negative limit as $\tau \to \infty$. Thus, the assertion follows by choosing v small. □

COROLLARY 7.4. *Let β belong to $\mathbb{R}$ and let $r_0 > 0$ be small and consider the following differential equation on $|r| < r_0$:*

$$\begin{aligned} \beta' &= -v\beta, \qquad r' = ru, \qquad \mathbf{s}' = \mathbf{w}, \\ u' &= \frac{u^2}{2} + \|\mathbf{w}\|^2 + V(\mathbf{s}) + r\tilde{f}(\beta), \\ \mathbf{w}' &= -\frac{u}{2}\mathbf{w} - \|\mathbf{w}\|^2\mathbf{s} - \nabla V(\mathbf{s}) + r\tilde{\mathbf{g}}(\beta). \end{aligned} \tag{7.4.1}$$

(1) *A point $p = (\beta, r, u, \mathbf{s}, \mathbf{w})^*$ is a rest point if and only if*

$$\beta^* = 0, \quad r^* = 0, \quad \mathbf{w}^* = 0, \quad (u^*)^2/2 = -V(\mathbf{s}^*), \quad \nabla V(\mathbf{s}^*) = 0. \tag{7.4.2}$$

(2) *Let J be the negative of the Hessian of V at a critical point. Then, the linearization of this vector field at a rest point is given by*

$$A = \begin{bmatrix} -v & 0 & 0 & 0 \\ 0 & u & 0 & 0 \\ 0 & 0 & u & \# \\ 0 & 0 & 0 & B \end{bmatrix}^*, \qquad B = \begin{pmatrix} 0 & I \\ -J & aI \end{pmatrix}^*, \qquad a^* = -\frac{u^*}{2}.$$

(3) *The eigenvalues of A are $-v$; u^*, with multiplicity two; and $\kappa^{\pm}(\lambda) = \frac{1}{2}(a \pm \sqrt{[a^2 - 4\lambda]})$ where λ is an eigenvalue of J.*

(4) *Thus, $\kappa^{\pm}(\lambda) = 0$ if and only if $\lambda = 0$. Also,* $\operatorname{Re}(\kappa^{\pm}(\lambda)) = 0$ *if and only if $\lambda = 0$.*

(5) *Consider a solution on which $r \to 0$ as $\tau \to \infty$. Such a solution goes to a rest point and u goes to a negative limit.*

Proof. The only point that needs discussion here is that $r = 0$ for any rest point. The rest is similar to what we have done in § 3. The potential function V has a negative maximum. Moreover, f is bounded and $\mathbf{w} = 0$ at a rest point. Thus, when r_0 is sufficiently small and $u = 0$, u' cannot vanish. Thus, at rest points $u \neq 0$ and $r = 0$. □

7.5. Important remarks.

(1) Consider a collinear central configuration $\mathcal{S}_0 = \{g\sigma \mid g \in \mathrm{SO}(3)\}$, where $\sigma = \mathbf{s}^*$ is a critical point of V that comes from a rest point p as in Corollary 7.4 above. We have seen that when considered as a quadratic form on $T_\sigma\Sigma$: $-J = HV(\sigma)$ vanishes on $\mathbf{X}_\sigma = T_\sigma(G \cdot \sigma)$ (Corollary 6.8(2)); is negative definite on $\mathbf{Y}_\sigma$ (6.10.2); and positive definite on $\mathbf{Z}_\sigma$ (Lemma 6.11). Since $\kappa^-(\lambda) = 0$ if and only if $\lambda = 0$, the matrix B, and hence A, has zero as an eigenvalue if and only if λ is zero.

(2) The kernel of B, that is, its zero eigenspace, is given by $K = \mathbf{X}_\sigma \times \{\mathbf{0}\}$. Since J is symmetric and its zero eigenspace is $\mathbf{X}_\sigma$, J, and hence $B \mid K$, can be diagonalized. (This is because if y is an eigenvector of J with eigenvalue λ, then $(y, \kappa y)^\dagger$ is an eigenvector of B. But $\lambda = 0$ if and only if $\kappa^- = 0$.) Moreover, the variables β, r, u, $\mathbf{s}$,

and $\mathbf{w}$ are orthogonal. When $u^* < 0$, the eigenvalues of A in the direction of the first three variables are negative. Now we let

$$\mathscr{S} = \{(\beta, r, u) \mid \beta = 0,\ r = 0,\ u = u^*\} \times \mathscr{S}_0 \times \{\mathbf{w} = \mathbf{0}\}. \tag{7.5.1}$$

The eigenspace of the zero eigenvalue of A is $T_p\mathscr{S}$. The real parts of the rest of the eigenvalues do not vanish. We remark that this analysis is true for any $p \in \mathscr{S}$ since any point in $\mathscr{S}_0$ is a critical point for V.

(3) Recall that $\mathscr{S}$ is an isolated compact manifold of rest points.

(4) Now we are in a position to apply the centre manifold theorem at each point in $\mathscr{S}$. Our main reference on this issue is § 9.2 of [Ch-H]. In the next two articles we will state simplified versions of their assertions, which are sufficient for our purpose.

Notation. Let $\mathbb{R}^{n+m} = \mathbb{R}^n \times \mathbb{R}^m$. We shall denote by $C^k(\mathbb{R}^{n+m}, \mathbb{R}^s)$ the set of C^k functions whose derivatives up to order r are uniformly bounded on $\mathbb{R}^{n+m}$, equipped with the C^k norm $|\cdot|_k$.

THEOREM 7.6 (Chow and Hale, [Ch-H]). *Let $(x, y) \in \mathbb{R}^{n+m}$, and let $u \in C^k(\mathbb{R}^{n+m}, \mathbb{R}^n)$ and $v \in C^k(\mathbb{R}^{n+m}, \mathbb{R}^m)$ satisfy $|u|_k, |v|_k < \varepsilon$, for some positive ε. Let C and D be constant matrices and consider the system*

$$\dot{x} = Cx + u(x, y), \qquad \dot{y} = Dy + v(x, y). \tag{7.6.1}$$

Assume that all eigenvalues of C have zero real parts, and all eigenvalues of D have nonzero real parts. Then, if ε is small enough, there is a function h in $C^k(\mathbb{R}^n, \mathbb{R}^m)$, which depends on u and v such that

(a) *$h(0) = 0$ when $u = v = 0$,*

(b) *The set $M_c = \{(x, y) \mid y = h(x)\}$, called the centre manifold, is a C^k manifold in $\mathbb{R}^{n+m}$ and is invariant under the flow of* (7.6.1);

(c) *M_c contains exactly those solutions $(x(t), y(t))$ for which $\sup_t |y(t)| < \infty$.*

Moreover, M_c is unique with respect to properties (a)-(c). □

THEOREM 7.7 (Chow and Hale, [Ch-H]). *There is a homeomorphism that takes the orbits of* (7.6.1) *to those of*

$$\dot{w} = Cw + u(w, h(w)), \qquad \dot{z} = Dz, \tag{7.7.1}$$

which sends the centre manifold M_c to $\{z = 0\}$, and preserves the sense of time. □

Theorem 7.7 is a generalization of the Hartman-Grobman theorem and means that we can linearize the flow in the normal direction to the centre manifold in a small neighbourhood of the rest point.

THEOREM 7.8. *Assume that $\mathscr{M}$ is an isolated compact regular submanifold that consists entirely of rest points of* (7.6.1).

(1) *For each $p \in \mathscr{M}$, there is a small neighbourhood U_p and a coordinate system (x, y) on U_p, in terms of which $U_p \cap \mathscr{M} = \{q \in U_p \mid y(q) = 0\}$, and the vector field* (7.6.1) *on U_p satisfies*

$$C = 0, \quad u(x, 0) = 0, \quad v(x, 0) = 0, \quad |u|_k, |v|_k < \varepsilon. \tag{7.8.1}$$

(2) *The centre manifold at p is $M_c(p) = U_p \cap \mathscr{M} = \{q \in U_p \mid y(q) = 0\}$, i.e., $h \equiv 0$.*

(3) *For a fixed but arbitrary U_p, let the coordinate system that gives the linearization* (7.7.1) *be $(w, z) : U_p \to \mathbb{R}^{n+m}$. Then* (7.7.1) *takes the form*

$$\dot{w} = 0, \qquad \dot{z} = Dz, \tag{7.8.2}$$

and $U_p \cap \mathscr{M} = \{q \in U_p \mid z(q) = 0\}$. Hence, U_p is foliated by $Z(p_0) = \{q \in U_p \mid w(q) = w(p_0)\}$, $p_0 \in U_p \cap \mathscr{M}$. Moreover, each $Z(p_0)$ is invariant and p_0 is a hyperbolic fixed point for the

flow on it. The local stable and unstable manifolds of p_0, relative to the flow on $Z(p_0)$, are the same ones that come from the whole flow.

(4) *$\mathcal{M}$ has a neighbourhood $\mathcal{U}$ which is foliated by the flow into invariant leaves Z_p, $p \in \mathcal{M}$, each of which intersects $\mathcal{M}$ exactly at p, and has p as a hyperbolic fixed point.*

Proof. (1) Since $\mathcal{M}$ is a regular submanifold, there is a coordinate system $\phi = (x, y)$ on a neighbourhood U_p of p such that, $y(q) = 0$ for all $q \in U_p \cap \mathcal{M}$, and $\phi(p) = 0$. We can choose U_p small enough so that we can apply Theorem 7.6. Since $\mathcal{M}$ is invariant, $v(x, 0) = 0$. Since it consists of rest points, $C = 0$ and $u(x, 0) = 0$. If we take U_p small enough, we get $|u|_k$, $|v|_k < \varepsilon$.

(2) Each point in $\mathcal{M} \cap U$ is a rest point. If $M_c(p)$ was not $U_p \cap \mathcal{M}$ and if $q \in U_p \cap \mathcal{M}$ then, by Theorem 7.6(c), the y-coordinate of the solution starting at q would be unbounded. But q is a rest point. Thus, $U_p \cap \mathcal{M} = M_c(p)$ and hence, $h(x) = 0$ for all x.

(3) The homeomorphism in Theorem 7.7 sends the centre manifold of (7.6.1) to that of the linearized flow. Thus, $U_p \cap \mathcal{M} = \{q \in U_p \mid z(q) = 0\}$. Since C and h vanish identically, and since $u(x, 0) = 0$, it follows that the flow on U_p is given by (7.8.2). Thus, each $Z(p_0)$ is invariant and the flow on it is given by the second equation above. Moreover, p_0 is a hyperbolic fixed point for the flow on $Z(p_0)$, and U_p is foliated as mentioned above. It is obvious that the local stable manifold of p_0 relative to this restricted flow coincides with the one that comes from the whole flow. The same is true for the unstable manifold.

(4) Since $\mathcal{M}$ is compact, it can be covered by a finite number of neighbourhoods of the type we saw in the previous paragraph. Let $p \in \mathcal{M}$ lie in the intersection of two such neighbourhoods. The two leaves that are attached to p and are obtained from the two different coordinate systems intersect at the stable and unstable manifolds of p, and hence must coincide, since in each neighbourhood each point lies in one and only one such a leaf. Thus, in the intersection, the foliation is well defined and the neighbourhood $\mathcal{U}$ can be any open neighbourhood contained in the union of the finitely many ones above. □

THEOREM 7.9. *Let $\phi(\tau)$ be a solution of* (7.6.1) *which approaches $\mathcal{M}$ as $\tau \to \infty$. Then, $\phi(\tau)$ converges to a single point in $\mathcal{M}$. In other words, the stable manifold of $\mathcal{M}$ is given by $St(\mathcal{M}) = \bigcup_{p \in \mathcal{M}} St(p)$. A similar statement holds for the unstable manifold of $\mathcal{M}$.*

Proof. Assume τ_0 is large enough for $\phi(\tau)$ to belong to $\mathcal{U}$ for all $\tau > \tau_0$. Hence, $\phi(\tau_0)$ belongs to one and only one leaf Z_p for some p in $\mathcal{M}$. If $\phi(\tau_0)$ was not on the stable manifold of p, $\phi(\tau)$ would have to leave $\mathcal{U}$ for some $\tau > \tau_0$, contradicting our assumption. Thus, $\phi(\tau)$ converges to p. □

COROLLARY 7.10. *A cluster μ that approaches a collinear central configuration does not undergo an infinite spin.*

Proof. Let $\mathcal{M} = \mathcal{S}$ in the previous two theorems, where $\mathcal{S}$ is given by (7.5.1). It is not hard to see that the vector field (7.4.1) can be written directly in the form (7.8.1) for as we mentioned in part (2) of Theorem 7.6, the matrix B is diagonalizable when restricted to its zero eigenspace. Actually, this part of B is what produces the matrix C in Theorem 7.6. Thus, C is diagonalizable, and the coordinate x can be chosen so that C is diagonal, and hence vanishes identically as all its eigenvalues are zeros. Since each point in $\mathcal{S} \cap U$ is a rest point, it follows that both $u(x, 0) = 0$ and $v(x, 0) = 0$. Also, if U is small enough, $|u|_k$, $|v|_k < \varepsilon$.

Acknowledgments. I express my gratitude to Professor Richard McGehee for introducing me to the subject of singularities in celestial mechanics and for his valuable guidance, time, and patience. I also thank Professor R. Moeckel for very helpful discussions.

REFERENCES

[A-M] R. ABRAHAM AND J. MARSDEN, *Foundations of Mechanics*, Benjamin-Cummings, Reading, MA, 1978.

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

[CC] C. CONLEY, *Isolated Invariant Sets and the Morse Index*, CBMS 38, American Mathematical Society, Providence, RI, 1978.

[Dv1] R. DEVANEY, *Collision orbits in the anisotropic Kepler problem*, Invent. Math., 45 (1978), pp. 221-251.

[Dv2] ———, *Triple collision in the planar isosceles three-body problem*, Invent. Math., 60 (1980), pp. 249-267.

[Dv3] ———, *Blowing up singularities in classical mechanical systems*, Amer. Math. Monthly, 89 (1982), pp. 535-552.

[McG1] R. MCGEHEE, *Triple collision in the collinear three-body problem*, Invent. Math., 27 (1974), pp. 191-227.

[McG2] ———, *Singularities in classical celestial mechanics*, in Proc. Internat. Congress of Mathematical, Helsinki, Finland, 1978.

[McG3] ———, *Von Zeipel's theorem on singularities in celestial mechanics*, Exposition. Math., 4 (1986), pp. 335-345.

[McG4] ———, *Lecture notes on ordinary differential equations*, University of Minnesota, Minneapolis, MN, 1981-1982.

[Moc1] R. MOECKEL, *Orbits near triple collision in the three-body problem*, Indiana Univ. Math. J., 32 (1983), pp. 221-240.

[Moc2] ———, *Heteroclinic phenomena in the isosceles three-body problem*, SIAM J. Math. Anal., 15 (1984), pp. 857-876.

[Moc3] ———, *Chaotic dynamics near triple collision*, Arch. Rational Mech. Anal., to appear.

[Pc] F. PACELLA, *Central configurations of the n-body problem via the equivariant Morse theory*, Arch. Rational Mech. Anal., 197 (1987), pp. 59-74.

[Pn] P. PAINLEVÉ, *Leçons sur la théorie analytique des équations différentielles*, Hermann, Paris, 1987.

[P-S] H. POLLARD AND D. G. SAARI, *Singularities of the n-body problem* I, Arch. Rational Mech. Anal., 30 (1968), pp. 263-269.

[S] D. SAARI, *The manifold structure for collision and for hyperbolic parabolic orbits in the n-body problem*, J. Differential Equations, 55 (1984), pp. 300-329.

[S1] ———, *Singularities and collisions of Newtonian gravitational systems*, Arch. Rational Mech. Anal., 49 (1973), pp. 311-320.

[S-H] D. SAARI AND N. HULKOWER, *On the manifold of total collapse orbits and of completely parabolic orbits for the n-body problem.* J. Differential Equations, 41 (1981), pp. 27-43.

[Sh] M. SHUB, *Diagonals and relative equilibria*, Manifolds-Amsterdam 1970, 199-201, Lecture Notes in Math., 197, Springer-Verlag, Berlin, 1971.

[Sp] H. SPERLING, *On the real singularities of the n-body problem*, J. Reine Angew. Math., 245 (1970), pp. 15-40.

[Su] K. SUNDMAN, *Recherches sur le problème des trois corps*, Acta Societatis Scientiarum Fennicae 34 (1906), pp. 1-43.

[W] A. WINTNER, *The Analytical Foundations of Celestial Mechanics*, Princeton University Press, Princeton, NJ, 1941.

[vZ] H. VON ZEIPEL, *Sur les singularités du problème des n Corps*, Ark. Mat., Astronomi och Fysik 4, 32 (1908), pp. 1-4.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1594-1618, November 1990
© 1990 Society for Industrial and Applied Mathematics
012

UNIFORM ASYMPTOTIC SOLUTIONS OF SECOND-ORDER LINEAR DIFFERENTIAL EQUATIONS HAVING A DOUBLE POLE WITH COMPLEX EXPONENT AND A COALESCING TURNING POINT*

T. M. DUNSTER†

Abstract. Second-order linear differential equations having a turning point and double pole with complex exponent are examined. The turning point is assumed to be a real continuous function of a parameter α, and coalesces with the pole at the origin when $\alpha \to 0$. Asymptotic expansions for solutions, as a second parameter $u \to \infty$, are constructed in terms of Bessel functions of purely imaginary order. The asymptotic solutions are uniformly valid for the argument lying in both real and complex regions that include both the coalescing turning point and the pole. The theory is then applied to obtain uniform asymptotic expansions for Legendre functions of large real degree and purely imaginary order.

Key words. asymptotic expansions, Legendre functions

AMS(MOS) subject classifications. 34E05, 34E20, 33A45

1. Introduction.

Consider the following second-order linear differential equation

$$\frac{d^2w}{dx^2} = \{u^2 f(\alpha, x) + g(\alpha, x)\}w, \tag{1.1}$$

where the parameters u and α are real, and the independent variable x lies in either a real interval (x_1, x_2) or a complex domain Δ, either of which may be unbounded. In a paper by Boyd and Dunster [1] asymptotic solutions of equations of the form (1.1) are derived for the case $u \to \infty$. In that paper the class of differential equation is that having a double pole, at $x = 0$, say, and a turning point (a zero of $f(\alpha, x)$) at $x = x_t(\alpha)$. The position of the turning point $x_t(\alpha)$ is assumed to be a continuous real function of α, and tended to zero as $\alpha \to 0$. Thus when $\alpha = 0$ the turning point coalesces with the pole at $x = 0$. It is supposed that there are no other critical points (poles or turning points) in the real or complex regions under consideration.

The asymptotic solutions derived in [1] are uniformly valid in a real interval, or a complex domain, that includes both critical points, and moreover the results are uniformly valid for $0 \leqq \alpha \leqq A$, for some positive constant A. In the case considered, the exponent of the pole is real, and the solutions are monotonic in a neighborhood of the pole. The asymptotic approximations for the solutions involve Bessel functions of large argument and variable real order $u\alpha$.

The purpose of the present paper is to tackle the complementary problem where, except when $\alpha = 0$, the exponent of the pole is purely imaginary, and thus all real solutions are oscillatory in a neighborhood of the pole. The essential difference between the previous results and the present is that in [1] the limit of $x^2 f(\alpha, x)$ as $x \to 0$ is assumed to be $\alpha^2/4$, whereas here the limit is assumed to be $-\alpha^2/4$. For this problem we will derive asymptotic solutions which involve Bessel functions of large argument and purely imaginary order $iu\alpha$, and our results will again be uniformly valid for $0 \leqq \alpha \leqq A$.

In a recent paper [4] new results have been recorded concerning Bessel functions of purely imaginary order, including asymptotic expansions, and identification of numerically satisfactory pairs. These results will be used in the subsequent analysis.

* Received by the editors August 28, 1989; accepted for publication (in revised form) December 11, 1989.
† Department of Mathematics, San Diego State University, San Diego, California 92182-0314.

Apart from the limit of $x^2 f(\alpha, x)$ being $-\alpha^2/4$ as $x \to 0$ all other assumptions on (1.1) are the same as in [1]. In the real variable case we suppose that $x^2 f(\alpha, x)$ and $x^2 g(\alpha, x)$ are infinitely differentiable[1] functions of x and continuous functions of x and α simultaneously. The limit of $x^2 g(\alpha, x)$ as $x \to 0$ is without loss of generality assumed to be equal to $-\frac{1}{4}$.

In the complex variable case $x = z$ lies in some domain D. Except at $z = 0$, the functions $f(\alpha, z)$ and $g(\alpha, z)$ will be assumed to be holomorphic functions of z lying in D and continuous functions of z and α simultaneously. As above we assume that the only critical points in D are a pole at $z = 0$ and a turning point on the real axis at $z = x_t(\alpha)$. The leading terms of the Laurent series of $f(\alpha, z)$ and $g(\alpha, z)$ are assumed to be $-\frac{1}{4}\alpha^2 z^{-2}$ and $-\frac{1}{4}z^{-2}$, respectively.

By means of a Liouville transformation we transform (1.1) into a new equation of the form

$$\frac{d^2 W}{d\zeta^2} = \left\{ u^2 \left(\frac{1}{4\zeta} - \frac{\alpha^2}{4\zeta^2} \right) + \frac{\psi(\alpha, \zeta)}{\zeta} - \frac{1}{4\zeta^2} \right\} W, \tag{1.2}$$

where the new independent variable $\zeta(x)$ is related to the original variable x by an integral equation. The variable ζ is real or complex according to whether x is real or complex. Equation (1.2) has the appropriate singular behavior at $\zeta = 0$, and a turning point at $\zeta = \alpha^2$; the pole and turning points of the original equation correspond, respectively, to the pole and turning points of the transformed equation.

The form of the function $\psi(\alpha, \zeta)/\zeta$ depends on the functions $f(\alpha, x)$ and $g(\alpha, x)$. If this term is neglected in (1.2) then the resulting equation, the so-called comparison equation, has exact solutions in terms of Bessel functions of order $iu\alpha$. From these Bessel functions, uniform asymptotic expansions will be constructed for solutions of (1.2), together with explicit error bounds. In §§ 2 and 3 we consider the real variable case, and in § 4 we consider the complex variable case.

There are two reasons why we consider the real variable case separately, rather than as merely a special case of complex variable theory. The first is that in the complex variable case the functions $f(\alpha, z)$ and $g(\alpha, z)$ must be (infinitely) differentiable functions of z (unequal to zero), whereas in the real variable case, $f(\alpha, x)$ and $g(\alpha, x)$ need only be finitely differentiable functions of x (unequal to zero). A second reason is that error bounds in the real variable case are generally simpler and sharper.

For differential equations having coalescing critical points, the first rigorous treatment involved two coalescing turning points. This problem has been tackled by Olver [7], who constructed uniform asymptotic solutions in terms of parabolic cylinder functions. Olver applied this general theory to derive uniform asymptotic approximations for Legendre functions [8] and Whittaker functions [9].

More recently Nestor [5] has derived uniform asymptotic solutions, in terms of Whittaker functions, for differential equations having a coalescing turning point and simple pole. Nestor considers the cases where solutions are monotonic, and those where they are oscillatory, near the pole, and applied the results to construct uniform approximations for Jacobi polynomials. Both Nestor and Olver have obtained explicit error bounds for their general asymptotic solutions.

The results of [1] have been used to construct uniform asymptotic expansions for Legendre functions, and subsequently have been used by Dunster to construct uniform asymptotic expansions for prolate spheroidal functions [2] and Whittaker functions

[1] This condition can be relaxed to that of finite differentiability if we require only a finite number of terms in the approximations, as opposed to asymptotic expansions.

[3]. The new results of this paper will be used in §§ 5 and 6 to construct asymptotic expansions for Legendre functions of large positive degree and purely imaginary order; § 5 deals with Legendre functions of real argument, and in § 6 Legendre functions of complex argument will be considered.

We plan to also apply the theory of the present paper to derive uniform asymptotic expansions for conical functions (Legendre functions of real order, and degree $-\frac{1}{2}+i\nu$, where ν is real), Whittaker functions, and oblate spheroidal functions.

Two remarks should perhaps be made concerning asymptotic solutions of differential equations having coalescing critical points. The first is that an application of the general theory to an equation can lead to approximations for functions that are uniformly valid for a wide range of the parameters involved. For example, in [1] asymptotic expansions are derived for Legendre functions of order μ and degree ν. These approximations are uniformly valid for large positive ν and for $0 \leqq \mu/(\nu+\frac{1}{2}) \leqq A$, where A is an arbitrary constant in $0<A<1$. Thus the results are uniformly valid for μ ranging from zero to $O(\nu)$.

The second remark is that it is highly desirable to obtain explicit error bounds. In [10] Olver mentions several advantages of having error bounds. In the case of coalescing critical points it is particularly important to have explicit error bounds because the uniform validity of the asymptotic theory may then be rigorously established (even if the bounds are sometimes difficult to compute numerically).

2. Formal series solutions and error bounds: Positive x and ζ. We denote the domain of (1.1) as the x interval (x_1, x_2), where x_1 is negative, x_2 is positive, and either may be infinite. We begin by considering real solutions of (1.1) in $0<x<x_2$, with the intention of constructing uniform asymptotic expansions for these solutions. In the real variable case we consider the intervals $x>0$ and $x<0$ separately, because generally no solution is real in both intervals.

The first step is to transform the original differential equation (1.1) to the new form (1.2). The standard method for doing this is by a Liouville transformation, which involves defining a new dependent and independent variable. For our problem this transformation is given as follows (cf. [1, Eq. (2.1)])

$$f(\alpha, x)\left(\frac{dx}{d\zeta}\right)^2 = \frac{1}{4\zeta} - \frac{\alpha^2}{4\zeta^2}, \tag{2.1a}$$

$$\left(\frac{d\zeta}{dx}\right)^{1/2} w(x) = W(\zeta). \tag{2.1b}$$

We integrate (2.1a) to obtain the following integral relationship between x and ζ:

$$\int_{\alpha^2}^{\zeta} \frac{(\xi-\alpha^2)^{1/2}}{2\xi}\, d\xi = \int_{x_t}^{x} \{f(\alpha, t)\}^{1/2}\, dt \qquad (x>x_t \text{ or } \zeta>\alpha^2), \tag{2.2a}$$

$$\int_{\alpha^2}^{\zeta} \frac{(\alpha^2-\xi)^{1/2}}{2\xi}\, d\xi = \int_{x_t}^{x} \frac{\{-t^2 f(\alpha, t)\}^{1/2}}{t}\, dt \qquad (x<x_t \text{ or } \zeta<\alpha^2). \tag{2.2b}$$

The lower integration limits are chosen to ensure that the turning point x_t of the original equation corresponds to the turning point α^2 of the transformed equation. This is essential in subsequent analysis. The endpoints $\zeta(x_1)$ and $\zeta(x_2)$ will be denoted by ζ_1 and ζ_2, respectively.

The square roots in the above integrals are taken to be nonnegative, and it is to be understood that the sign of ζ is to be the same as that of x. Thus in this section we shall construct asymptotic solutions for (2.2) in the interval $\zeta>0$.

Explicit integration of the left-hand sides of (2.2a, b) can be achieved (see [1, Eq. (2.3)]); we have, respectively,

$$(\zeta-\alpha^2)^{1/2}-\alpha\arctan\left\{\frac{(\zeta-\alpha^2)^{1/2}}{\alpha}\right\} \qquad (\zeta>\alpha^2), \tag{2.3a}$$

$$(\alpha^2-\zeta)^{1/2}-\frac{\alpha}{2}\ln\left\{\frac{\alpha+(\alpha^2-\zeta)^{1/2}}{|\alpha-(\alpha^2-\zeta)^{1/2}|}\right\} \qquad (\zeta<\alpha^2). \tag{2.3b}$$

When $\alpha=0$ the limiting form of (2.2a, b) applies, namely

$$\begin{aligned}\zeta^{1/2}&=\int_0^x\{f(0,t)\}^{1/2}\,dt \qquad (x>0 \text{ or } \zeta>0),\\ \{-\zeta\}^{1/2}&=\int_0^x\{-f(0,t)\}^{1/2}\,dt \qquad (x<0 \text{ or } \zeta<0).\end{aligned} \tag{2.4}$$

Whether or not $\alpha=0$ the pole $x=0$ is mapped to $\zeta=0$. By considering separately the intervals $(x_1,0)$, $(0,x_t)$, (x_t,x_2), and neighborhoods of $x=0$ and $x=x_t$, we can show that the function $\zeta(x)$ is monotonically increasing and infinitely differentiable in the interval $x_1<x<x_2$.

The effect of the Liouville transformation is to convert (1.1) to the form (1.2). For this particular transformation the Schwarzian function $\psi(\alpha,\zeta)$ is found to be

$$\psi(\alpha,\zeta)=\frac{\zeta+4\alpha^2}{16(\zeta-\alpha^2)^2}+\frac{(\zeta-\alpha^2)(4f\ddot{f}-5\dot{f}^2+16gf^2)}{64\zeta f^3}. \tag{2.5}$$

This function is infinitely differentiable in $\zeta_1<\zeta<\zeta_2$. In particular, it is so at $\zeta=0$, as can be seen by considering its Maclaurin series. We find that this series exists to all orders. It can be shown under certain conditions on $f(\alpha,x)$ that $\psi(\alpha,\zeta)$ is uniformly continuous at $\alpha=0$, $\zeta=0$ in the (α,ζ)-plane; see Lemma 1 of [1]. The lemma applies in the present circumstances with $p(\alpha,x)$ defined by

$$f(\alpha,x)=\frac{x-x_t(\alpha)}{4x^2}\,p(\alpha,x). \tag{2.6}$$

Let us now proceed to construct formal series solutions to (1.2). To do so we first observe that if the term $\psi(\alpha,\zeta)/\zeta$ is neglected in that equation, then the resulting equation

$$\frac{d^2W}{d\zeta^2}=\left\{u^2\left(\frac{1}{4\zeta}-\frac{\alpha^2}{4\zeta^2}\right)-\frac{1}{4\zeta^2}\right\}W \tag{2.7}$$

has exact solutions of the form $\zeta^{1/2}\mathscr{L}_{iu\alpha}(u\zeta^{1/2})$ when $\zeta>0$, where $\mathscr{L}$ denotes the modified Bessel functions K, I, or any linear combination of the two. For our purposes it is necessary to select a numerically satisfactory pair for $\zeta>0$ (see [6, p. 154]). It should be mentioned that $I_{i\nu}(x)$ is complex when ν and x are both positive, and as such is not an appropriate choice. On the other hand, $K_{i\nu}(x)$ is an appropriate choice, since it is real in the same circumstances, and moreover is recessive at $x=\infty$. In the corresponding problem of [4, § 7] the choice of real numerically satisfactory solutions is $K_{i\nu}$ and $L_{i\nu}$, where

$$L_{i\nu}(x)=\frac{\pi}{2\sinh(\nu\pi)}\{I_{i\nu}(x)+I_{-i\nu}(x)\} \qquad (x>0).$$

The function $L_{i\nu}$ is a numerically satisfactory companion to $K_{i\nu}(x)$ in $0<x<\infty$ for each *fixed positive* value of ν. However, the function is not defined when $\nu=0$. In the present case we require a numerically satisfactory companion to $K_{i\nu}(x)$ for $0\leqq\nu<\infty$. With these considerations in mind we define the following real solution as our companion to $K_{i\nu}(x)$:

$$\tilde{I}_{i\nu}(x)=2\sinh(\nu\pi)\,e^{-\nu\pi}L_{i\nu}(x)=\pi\,e^{-\nu\pi}\{I_{i\nu}(x)+I_{-i\nu}(x)\}. \tag{2.8}$$

Clearly this is defined for $0\leqq\nu<\infty$, and in particular $\tilde{I}_0(x)=2\pi I_0(x)$. Thus when $\nu=0$ $\tilde{I}_{i\nu}(x)$ has the characteristic property of being recessive at $x=0$.

Note the following asymptotic forms as $\nu\to\infty$ in terms of Airy functions (see [4, § 4]):

$$K_{i\nu}(\nu z)\sim\frac{\pi\,e^{-\nu\pi/2}}{\nu^{1/3}}\left(\frac{4\hat{\zeta}}{1-z^2}\right)^{1/4}\mathrm{Ai}\,(-\nu^{2/3}\hat{\zeta}), \tag{2.9}$$

$$\tilde{I}_{i\nu}(\nu z)\sim\frac{\pi\,e^{-\nu\pi/2}}{\nu^{1/3}}\left(\frac{4\hat{\zeta}}{1-z^2}\right)^{1/4}\mathrm{Bi}\,(-\nu^{2/3}\hat{\zeta}), \tag{2.10}$$

where

$$\frac{2}{3}\hat{\zeta}^{3/2}(z)=\ln\left\{\frac{1+(1-z^2)^{1/2}}{z}\right\}-(1-z^2)^{1/2}. \tag{2.11}$$

These asymptotic formulas are uniformly valid for $|\arg(z)|\leqq\pi-\delta(<\pi)$ in the complex z plane. The branches in (2.11) take their principal values when $z\in(0,1)$ and $\hat{\zeta}\in(0,\infty)$, and are continuous elsewhere. The points $z=0,\ 1,\ \infty$ on the real axis correspond to $\hat{\zeta}=\infty,\ 0,\ -\infty$, respectively. For a detailed discussion on this transformation see [6, pp. 419–422]; we have written Olver's variable ζ with a circumflex ($\hat{}$) here to avoid confusion with the different variable ζ defined by (2.2a, b).

From the asymptotic forms (2.9) and (2.10), and the known asymptotic behavior of Airy functions (see, e.g., [6, pp. 392–393]), we see that the functions $\tilde{I}_{i\nu}(x)$ and $K_{i\nu}(x)$ are oscillatory in $0<x<x_t(\alpha)$ with a phase difference of $\pi/2$, and exponentially large and small, respectively, in the interval $x_t(\alpha)<x<\infty$. It is these properties that make them a numerically satisfactory pair. For other properties of $K_{i\nu}(x)$ and $\tilde{I}_{i\nu}(x)$ see [4, § 2].

Having selected a numerically satisfactory pair of solutions for the comparison equation (2.7), we now seek formal series solutions for the full equation (1.2) of the form

$$\zeta^{1/2}\mathscr{L}_{iu\alpha}(u\zeta^{1/2})\sum_{s=0}^{\infty}\frac{A_s(\alpha,\zeta)}{u^{2s}}+\frac{\zeta}{u}\mathscr{L}'_{iu\alpha}(u\zeta^{1/2})\sum_{s=0}^{\infty}\frac{B_s(\alpha,\zeta)}{u^{2s}}\qquad(\zeta>0), \tag{2.12}$$

where $\mathscr{L}$ denotes K or $\tilde{I}$, or any combination of the two. Primes (′) denote the derivative with respect to the argument. After formally substituting (2.12) into (1.2) and following the same technique as in [1, § 2] we obtain the relations

$$B_s(\alpha,\zeta)=|\zeta-\alpha^2|^{-1/2}\int_{\alpha^2}^{\zeta}|\xi-\alpha^2|^{-1/2}\{\psi(\alpha,\xi)A_s(\alpha,\xi)-A'_s(\alpha,\xi)-\xi A''_s(\alpha,\xi)\}\,d\xi, \tag{2.13}$$

$$A_s(\alpha,\zeta)=-\zeta B'_{s-1}(\alpha,\zeta)+\int_{\alpha^2}^{\zeta}\psi(\alpha,\xi)B_{s-1}(\alpha,\xi)\,d\xi+\lambda_s. \tag{2.14}$$

In (2.14), $\{\lambda_s\}$ are arbitrary constants of integration. The corresponding integration constants in (2.13) are set to zero to ensure that the coefficient functions $\{B_s\}$ are differentiable at the turning point $\zeta=\alpha^2$. We choose $\lambda_0=A_0=1$; the remaining constants $\lambda_1,\ \lambda_2,\cdots$ can be chosen according to the particular application. Relations

(2.13) and (2.14) successively determine $B_0, A_1, B_1, \cdots$, and each of these is readily shown to be an infinitely differentiable function of ζ in the interval $0 \leqq \zeta < \zeta_2$ for each value of α in the interval $0 \leqq \alpha \leqq A$ (cf. [6, p. 410]).

The foregoing analysis is formal. Our task now is to terminate the series (2.12) after a finite number of terms, and then find an upper bound on the error. In doing this we will establish rigorously that the formal series solution is an asymptotic expansion of an exact solution of (1.2), uniformly valid for $0 < \zeta < \zeta_2$, $0 \leqq \alpha \leqq A$, $u > 0$.

Therefore we define the following expansion to be an exact solution of (1.2):

$$W_{2n+1}(u, \alpha, \zeta) = \zeta^{1/2}\mathscr{L}_{iu\alpha}(u\zeta^{1/2}) \sum_{s=0}^{n} \frac{A_s(\alpha, \zeta)}{u^{2s}} + \frac{\zeta}{u}\mathscr{L}'_{iu\alpha}(u\zeta^{1/2}) \cdot \sum_{s=0}^{n-1} \frac{B_s(\alpha, \zeta)}{u^{2s}} + \varepsilon_{2n+1}(u, \alpha, \zeta). \tag{2.15}$$

We obtain a bound for the unknown error term ε_{2n+1} by the following standard method. First, from the condition that W_{2n+1} satisfies the differential equation (1.2), we find that ε_{2n+1} satisfies an inhomogeneous differential equation (cf. [1, § 3]). Choosing an arbitrary value $\tilde{\zeta}$ of ζ in the interval $[0, \zeta_2)$ and applying the method of variation of parameters on the equation yields

$$\varepsilon_{2n+1}(u, \alpha, \zeta) = \int_{\tilde{\zeta}}^{\zeta} K(\zeta, \xi)\left[\frac{1}{u^{2n}}\{|\xi - \alpha^2|^{1/2}B_n(\alpha, \xi)\}'\xi^{1/2}\mathscr{L}_{iu\alpha}(u\zeta^{1/2}) + |\xi - \alpha^2|^{-1/2}\psi(\alpha, \xi)\varepsilon_{2n+1}(u, \alpha, \xi)\right] d\xi, \tag{2.16}$$

this being a Volterra integral equation for a solution that satisfies the boundary condition $\varepsilon_{2n+1}(u, \alpha, \tilde{\zeta}) = \varepsilon'_{2n+1}(u, \alpha, \tilde{\zeta}) = 0$. In this equation $K(\zeta, \xi)$ is given by

$$K(\zeta, \xi) = \frac{e^{\nu\pi}}{\pi}|\xi - \alpha^2|^{1/2}\{\zeta^{1/2}\tilde{I}_{iu\alpha}(u\zeta^{1/2})\xi^{-1/2}K_{iu\alpha}(u\xi^{1/2}) - \zeta^{1/2}K_{iu\alpha}(u\zeta^{1/2})\xi^{-1/2}\tilde{I}_{iu\alpha}(u\xi^{1/2})\}. \tag{2.17}$$

We shall use Theorem 10.1 of [6, Chap. 10] on the integral equation (2.16) to derive the desired bounds on ε_{2n+1} and ε'_{2n+1}. The essential part in the application of this theorem is to meet condition (v) with suitable continuous functions P_0, P_1, Q. The standard method to achieve this is to introduce so-called auxiliary functions, satisfying the relations

$$\tilde{I}_{i\nu}(x) = E_{\nu,1}(x)M^+_\nu(x)\cos\{\theta^+_\nu(x)\}, \tag{2.18}$$

$$K_{i\nu}(x) = E_{\nu,2}(x)M^+_\nu(x)\sin\{\theta^+_\nu(x)\}. \tag{2.19}$$

The weight functions $E_{\nu,1}(x)$ and $E_{\nu,2}(x)$ are to be continuous monotonic functions of x. They must be prescribed in such a way that Theorem 10.1 of Olver can be applied, such that the resulting error bounds reflect the asymptotic behavior of the error term. Once chosen, the modulus and phase auxiliary functions $M^+_\nu(x)$ and $\theta^+_\nu(x)$ are given implicitly by the relations (2.18) and (2.19).

We define the weight functions as follows. Let $x = X_\nu$ be the largest positive root of

$$K_{i\nu}(x) - \tilde{I}_{i\nu}(x) = 0. \tag{2.20}$$

It can be shown that $X_0 > 0$, and that X_ν is a monotonically increasing function of ν. Moreover, for each positive value of ν

$$0 < k_{\nu,1} < l_{\nu,1} < X_\nu < \infty,$$

where $k_{\nu,1}$ and $l_{\nu,1}$ are, respectively, the largest zeros of $K_{i\nu}(x)$ and $\tilde{I}_{i\nu}(x)$ (cf. [4, eq. (6.2)]).

Let $x = c$ denote the negative root of the equation $\mathrm{Ai}(x) = \mathrm{Bi}(x)$ ([6, p. 395]). Then from (2.9)-(2.11) we find that as $\nu \to \infty$

$$X_\nu = \nu + c(\tfrac{1}{2}\nu)^{1/3} + O(\nu^{-1/3}). \tag{2.21}$$

We prescribe

$$E_{\nu,1}(x) = \begin{cases} 1 & (0 < x \leqq X_\nu), \\ (\tilde{I}_{i\nu}(x)/K_{i\nu}(x))^{1/2} & (X_\nu \leqq x < \infty), \end{cases} \tag{2.22}$$

$$E_{\nu,2}(x) = \begin{cases} K_{i\nu}(X_\nu)^{-1} \sup_{x \leqq \xi \leqq X_\nu} |K_{i\nu}(\xi)| & (0 < x \leqq X_\nu), \\ (K_{i\nu}(x)/\tilde{I}_{i\nu}(x))^{1/2} & (X_\nu \leqq x < \infty). \end{cases} \tag{2.23}$$

In defining $E_{\nu,2}(x)$ we have taken into account the behavior of $K_{i\nu}(x)$ near $x = 0$; see [4, Eq. (2.14)]. In particular, it should be noted that the oscillation amplitude of the function near $x = 1$ becomes unbounded as $\nu \to 0$, reflecting its logarithmic behavior at the singularity when $\nu = 0$. The factor $K_{i\nu}(X_\nu)^{-1}$ is introduced to ensure continuity of the weight function at $x = X_\nu$.

The weight functions $E_{\nu,1}(x)$ and $E_{\nu,2}(x)$ are, respectively, monotonically increasing and decreasing functions of x for $0 < x < \infty$ (cf. [4, eq. (6.4)]). As $x \to \infty$ we find that

$$E_{\nu,1}(x),\ E_{\nu,2}(x)^{-1} \sim \sqrt{2}\, e^{-\nu\pi/2}\, e^{x}. \tag{2.24}$$

Having defined these weight functions we find from (2.18) and (2.19) that the modulus and phase functions are real, and are given by

$$M_\nu^+(x) = \begin{cases} \{\tilde{I}_{i\nu}(x)^2 + K_{i\nu}(x)^2/E_{\nu,2}(x)^2\}^{1/2} & (0 < x \leqq X_\nu), \\ \{2\tilde{I}_{i\nu}(x)K_{i\nu}(x)\}^{1/2} & (X_\nu \leqq x < \infty), \end{cases} \tag{2.25}$$

$$\theta_\nu^+(x) = \begin{cases} \arctan\left\{\dfrac{K_{i\nu}(x)}{E_{\nu,2}(x)\tilde{I}_{i\nu}(x)}\right\} & (0 < x \leqq X_\nu), \\ \dfrac{\pi}{4} & (X_\nu \leqq x < \infty). \end{cases} \tag{2.26}$$

The branch on the inverse tangent is chosen so that $\theta_\nu^+(x)$ is continuous. Note that as $x \to \infty$

$$M_\nu^+(x) \sim \left(\frac{2\pi\, e^{-\nu\pi}}{x}\right)^{1/2}, \tag{2.27}$$

and when $\nu = 0$ and $x \to 0$

$$E_{0,2}(x) \sim \frac{|\ln(x)|}{K_0(X_0)}, \tag{2.28a}$$

$$M_0^+(x) \to (4\pi^2 + K_0^2(X_0))^{1/2}. \tag{2.28b}$$

Next, we introduce modulus and phase functions for the derivatives. We define them implicitly as the real functions satisfying

$$\tilde{I}'_{i\nu}(x) = E_{\nu,1}(x) N_\nu^+(x) \cos\{\omega_\nu^+(x)\}, \tag{2.29}$$

$$K'_{i\nu}(x) = E_{\nu,2}(x) N_\nu^+(x) \sin\{\omega_\nu^+(x)\}. \tag{2.30}$$

The following constants will appear in the subsequent error bounds:

$$\kappa^+ = \sup\left\{\frac{e^{\nu\pi}}{\pi}|x^2-\nu^2|^{1/2}E_{\nu,1}(x)E_{\nu,2}(x)\{M_\nu^+(x)\}^2\right\}, \tag{2.31}$$

$$\mu_1^+ = \sup\left\{\frac{e^{\nu\pi}}{\pi}|x^2-\nu^2|^{1/2}E_{\nu,2}(x)M_\nu^+(x)|\tilde{I}_{i\nu}(x)|\right\}, \tag{2.32}$$

$$\mu_2^+ = \sup\left\{\frac{e^{\nu\pi}}{\pi}|x^2-\nu^2|^{1/2}E_{\nu,1}(x)M_\nu^+(x)|K_{i\nu}(x)|\right\}, \tag{2.33}$$

each supremum being evaluated over $x>0$ and $\nu\geqq 0$. The existence of these suprema, and the corresponding ones for the negative variable case (see (3.12)-(3.14) below), can be established in a manner similar to the proof of Lemma 2 in [1, App. B], using the asymptotic results of [4, §§ 2-4].

We now are in a position to state our theorem on error bounds, the proof of which has been outlined above.

THEOREM 1. *With the conditions given in this and the previous section,* (1.2) *has, for each positive* u, *nonnegative* α, *and nonnegative integer* n, *the following pair of solutions, which are infinitely differentiable in* $0<\zeta<\zeta_2$:

$$W_{2n+1,1}(u,\alpha,\zeta) = \zeta^{1/2}\tilde{I}_{iu\alpha}(u\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{u^{2s}} + \frac{\zeta}{u}\tilde{I}'_{iu\alpha}(u\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{u^{2s}} + \varepsilon_{2n+1,1}(u,\alpha,\zeta), \tag{2.34}$$

$$W_{2n+1,2}(u,\alpha,\zeta) = \zeta^{1/2}K_{iu\alpha}(u\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{u^{2s}} + \frac{\zeta}{u}K'_{iu\alpha}(u\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{u^{2s}} + \varepsilon_{2n+1,2}(u,\alpha,\zeta), \tag{2.35}$$

where

$$\frac{|\varepsilon_{2n+1,1}(u,\alpha,\zeta)|}{\zeta^{1/2}M_{u\alpha}^+(u\zeta^{1/2})}, \frac{|\partial\varepsilon_{2n+1,1}(u,\alpha,\zeta)/\partial\zeta|}{\frac{1}{2}uN_{u\alpha}^+(u\zeta^{1/2})+\frac{1}{2}\zeta^{-1/2}M_{u\alpha}^+(u\zeta^{1/2})}$$

$$\leqq\frac{\mu_1^+}{u^{2n+1}}E_{u\alpha,1}(u\zeta^{1/2})\mathscr{V}_{0,\zeta}\{|\zeta-\alpha^2|^{1/2}B_n(\alpha,\zeta)\} \cdot\exp\left\{\frac{\kappa^+}{u}\mathscr{V}_{0,\zeta}\{|\zeta-\alpha^2|^{1/2}B_0(\alpha,\zeta)\}\right\}, \tag{2.36}$$

$$\frac{|\varepsilon_{2n+1,2}(u,\alpha,\zeta)|}{\zeta^{1/2}M_{u\alpha}^+(u\zeta^{1/2})}, \frac{|\partial\varepsilon_{2n+1,2}(u,\alpha,\zeta)/\partial\zeta|}{\frac{1}{2}uN_{u\alpha}^+(u\zeta^{1/2})+\frac{1}{2}\zeta^{-1/2}M_{u\alpha}^+(u\zeta^{1/2})}$$

$$\leqq\frac{\mu_2^+}{u^{2n+1}}E_{u\alpha,2}(u\zeta^{1/2})\mathscr{V}_{\zeta,\zeta_2}\{|\zeta-\alpha^2|^{1/2}B_n(\alpha,\zeta)\} \cdot\exp\left\{\frac{\kappa^+}{u}\mathscr{V}_{\zeta,\zeta_2}\{|\zeta-\alpha^2|^{1/2}B_0(\alpha,\zeta)\}\right\}. \tag{2.37}$$

When $\zeta_2=\infty$, (2.37) is meaningful, and the expansions (2.34), (2.35) are uniformly valid throughout $0<\zeta<\infty$, provided the variations of $(\zeta-\alpha^2)^{1/2}B_s(\zeta)$ $(s=0, 1, 2, \cdots)$

converge at infinity. A sufficient condition for this to be true is for the ζ-derivatives to satisfy

$$\psi^{(s)}(\alpha, \zeta) = O(\zeta^{-1/2-s-\sigma}), \qquad \zeta \to \infty, \tag{2.38}$$

where σ is any positive constant (cf. [6, p. 445, Ex. 4.2]).

The importance of the bounds (2.36), (2.37) is that from them we can establish the asymptotic nature of the error terms, and hence also of the expansions (2.12). For example, consider the behavior of $\varepsilon_{2n+1,1}$ as $\zeta \to 0$. From (2.36) we see that

$$\varepsilon_{2n+1,1}(u, \alpha, \zeta) = \zeta^{1/2} \tilde{I}_{iu\alpha}(u\zeta^{1/2}) O(\zeta) \quad \text{as } \zeta \to 0. \tag{2.39}$$

Consider next the asymptotic behavior of $\varepsilon_{2n+1,1}$ as $u \to \infty$. The bound (2.38) can be used to show that (2.12), with $\mathscr{L} = \tilde{I}$, is a uniformly valid compound asymptotic expansion of $W_{2n+1,1}(u, \alpha, \zeta)$. This can be achieved in a similar manner to that in [1, § 3]. Discussions similar to those above hold for the second solution $W_{2n+1,2}$.

Finally, we remark that there are solutions $W_1(u, \alpha, \zeta)$ and $W_2(u, \alpha, \zeta)$, independent of n, which have the infinite series (2.12), with $\mathscr{L} = \tilde{I}$, K, respectively, as their compound asymptotic expansions. This can be shown in a manner similar to that of [6, Chap. 10, § 6].

3. Formal series solutions and error bounds: Negative x and ζ. Now consider (2.7) for the case $\zeta < 0$. Exact numerically satisfactory solutions of this comparison equation are $|\zeta|^{1/2} \mathscr{C}_{iu\alpha}(u|\zeta|^{1/2})$, where $\mathscr{C}_{i\nu}$ denotes either of the Bessel functions $F_{i\nu}$ and $G_{i\nu}$, which are defined in [4, § 3]. Relevant properties of these functions are given in this reference. An important observation is that both functions are oscillatory in $0 < x < \infty$ for $\nu > 0$, and when $\nu = 0$, $G_0(x)$ is logarithmically singular at $x = 0$, and $F_0(x)$ is bounded at $x = 0$.

For correspondence to (2.12) we seek formal series solutions to (1.2) for $\zeta < 0$ of the form

$$|\zeta|^{1/2} \mathscr{C}_{iu\alpha}(u|\zeta|^{1/2}) \sum_{s=0}^{\infty} \frac{A_s(\alpha, \zeta)}{u^{2s}} + \frac{|\zeta|}{u} \mathscr{C}'_{iu\alpha}(u|\zeta|^{1/2}) \sum_{s=0}^{\infty} \frac{B_s(\alpha, \zeta)}{u^{2s}} \qquad (\zeta < 0), \tag{3.1}$$

and in doing so we find that the coefficients $\{A_s\}$ and $\{B_s\}$ are the same as those given by (2.13) and (2.14) for the case $\zeta > 0$. It is readily shown that these coefficients are infinitely differentiable in $\zeta_1 < \zeta < \zeta_2$.

Before stating our theorem on error bounds we define auxiliary functions as follows. We define a continuous monotonically decreasing weight function for $G_{i\nu}(x)$ by

$$E_{\nu,4}(x) = \max\{1, \sup_{x \leqq \xi < \infty} |G_{i\nu}(\xi)|\} \qquad (x > 0). \tag{3.2}$$

The introduction of this weight function is necessary on account of the unbounded behavior of $G_{i\nu}(x)$ near $x = 0$ as $\nu \to 0$ (cf. (2.23) above). Note that for every nonnegative ν, $E_{\nu,4}(x) = 1$ for sufficiently large x. Also, there exists a positive value of ν, ν_1 say, such that $E_{\nu,4}(x) \equiv 1$ identically for all $\nu \geqq \nu_1$ (see [4, Eq. (5.16)]).

We next set

$$F_{i\nu}(x) = M_\nu^-(x) \cos\{\theta_\nu^-(x)\}, \tag{3.3}$$

$$G_{i\nu}(x) = E_{\nu,4}(x) M_\nu^-(x) \sin\{\theta_\nu^-(x)\}. \tag{3.4}$$

The modulus and phase functions are real functions explicitly given by

$$M_\nu^-(x) = \{F_{i\nu}(x)^2 + G_{i\nu}(x)^2 / E_{\nu,4}(x)^2\}^{1/2}, \tag{3.5}$$

$$\theta_\nu^-(x) = \arctan\left\{\frac{G_{i\nu}(x)}{E_{\nu,4}(x) F_{i\nu}(x)}\right\}. \tag{3.6}$$

The inverse tangent is chosen to ensure that $\theta_\nu^-(x)$ is continuous, and

$$\theta_\nu^-(x) = x - \frac{\pi}{4} + o(1) \quad \text{as } x \to \infty. \tag{3.7}$$

Other asymptotic behaviors are as follows. As $x \to \infty$

$$M_\nu^-(x) \sim \left(\frac{2}{\pi x}\right)^{1/2}, \tag{3.8}$$

and when $\nu = 0$, $x \to 0$,

$$E_{0,4}(x) \sim \frac{2}{\pi} |\ln(x)|, \tag{3.9a}$$

$$M_0^-(x) \to \sqrt{2}. \tag{3.9b}$$

For the derivatives we define

$$F'_{i\nu}(x) = N_\nu^-(x) \cos\{\omega_\nu^-(x)\}, \tag{3.10}$$

$$G'_{i\nu}(x) = E_{\nu,4}(x) N_\nu^-(x) \sin\{\omega_\nu^-(x)\}. \tag{3.11}$$

The following constants will appear in the theorem on error bounds:

$$\kappa^- = \sup\{\pi(x^2+\nu^2)^{1/2} E_{\nu,4}(x)\{M_\nu^-(x)\}^2\}, \tag{3.12}$$

$$\mu_1^- = \sup\{\pi(x^2+\nu^2)^{1/2} E_{\nu,4}(x) M_\nu^-(x) |F_{i\nu}(x)|\}, \tag{3.13}$$

$$\mu_2^- = \sup\{\pi(x^2+\nu^2)^{1/2} M_\nu^-(x) |G_{i\nu}(x)|\}. \tag{3.14}$$

Again, these suprema are evaluated over $x > 0$ and $\nu \geqq 0$.

We now apply Theorem 10.1 of [6, Chap. 6] to obtain the following result.

THEOREM 2. *With the conditions given in* §§ 1 *and* 2, (1.2) *has, for each positive* u, *nonnegative* α, *and nonnegative integer* n, *the following pair of solutions, which are infinitely differentiable in* $\zeta_1 < \zeta < 0$:

$$\begin{aligned} W_{2n+1,3}(u, \alpha, \zeta) &= |\zeta|^{1/2} F_{iu\alpha}(u|\zeta|^{1/2}) \sum_{s=0}^{n} \frac{A_s(\alpha, \zeta)}{u^{2s}} \\ &\quad + \frac{|\zeta|}{u} F'_{iu\alpha}(u|\zeta|^{1/2}) \sum_{s=0}^{n-1} \frac{B_s(\alpha, \zeta)}{u^{2s}} + \varepsilon_{2n+1,3}(u, \alpha, \zeta), \end{aligned} \tag{3.15}$$

$$\begin{aligned} W_{2n+1,4}(u, \alpha, \zeta) &= |\zeta|^{1/2} G_{iu\alpha}(u|\zeta|^{1/2}) \sum_{s=0}^{n} \frac{A_s(\alpha, \zeta)}{u^{2s}} \\ &\quad + \frac{|\zeta|}{u} G'_{iu\alpha}(u|\zeta|^{1/2}) \sum_{s=0}^{n-1} \frac{B_s(\alpha, \zeta)}{u^{2s}} + \varepsilon_{2n+1,4}(u, \alpha, \zeta), \end{aligned} \tag{3.16}$$

where

$$\begin{aligned} &\frac{|\varepsilon_{2n+1,3}(u, \alpha, \zeta)|}{|\zeta|^{1/2} M_{u\alpha}^-(u|\zeta|^{1/2})}, \frac{|\partial \varepsilon_{2n+1,3}(u, \alpha, \zeta)/\partial\zeta|}{\frac{1}{2}u N_{u\alpha}^-(u|\zeta|^{1/2}) + \frac{1}{2}|\zeta|^{-1/2} M_{u\alpha}^-(u|\zeta|^{1/2})} \\ &\leqq \frac{\mu_1^-}{u^{2n+1}} \mathcal{V}_{\zeta,0}\{|\zeta - \alpha^2|^{1/2} B_n(\alpha, \zeta)\} \\ &\quad \cdot \exp\left\{\frac{\kappa^-}{u} \mathcal{V}_{\zeta,0}\{|\zeta - \alpha^2|^{1/2} B_0(\alpha, \zeta)\}\right\}, \end{aligned} \tag{3.17}$$

$$\frac{|\varepsilon_{2n+1,4}(u,\alpha,\zeta)|}{|\zeta|^{1/2}M^{-}_{u\alpha}(u|\zeta|^{1/2})},\ \frac{|\partial\varepsilon_{2n+1,4}(u,\alpha,\zeta)/\partial\zeta|}{\frac{1}{2}uN^{-}_{u\alpha}(u|\zeta|^{1/2})+\frac{1}{2}|\zeta|^{-1/2}M^{-}_{u\alpha}(u|\zeta|^{1/2})}$$

$$\leqq\frac{\mu_2^-}{u^{2n+1}}E_{u\alpha,4}(u|\zeta|^{1/2})\mathscr{V}_{\zeta_1,\zeta}\{|\zeta-\alpha^2|^{1/2}B_n(\alpha,\zeta)\} \tag{3.18}$$

$$\cdot\exp\left\{\frac{\kappa^-}{u}\mathscr{V}_{\zeta_1,\zeta}\{|\zeta-\alpha^2|^{1/2}B_0(\alpha,\zeta)\}\right\}.$$

If as $\zeta\to\infty$ the ζ-derivatives $\psi^{(s)}(\alpha,\zeta)$ are $O(|\zeta|^{-1/2-s-\sigma})$ for some positive σ, then ζ_1 can be set to $-\infty$, and the expansions will be uniformly valid in the ζ-interval $(-\infty,0)$.

The bounds (3.17), (3.18) can be used to deduce the asymptotic behavior, with respect to both ζ and u, of expansions (3.1). The discussions are similar to the corresponding ones following Theorem 1.

4. Uniform asymptotic expansions: Complex z and ζ. Having tackled the real variable case, let us now consider the case where the variables x (hereafter denoted by z) and ζ are complex. Our task is to derive asymptotic expansions for solutions of (1.1), via (1.2), with ζ now lying in some complex domain Δ containing both the critical points $\zeta=0$ and $\zeta=\alpha^2$.

The Liouville transformation equations (2.1a, b) still apply (with x replaced by z), transforming (1.1) to the new form (1.2). Equations (2.2a, b) are of course no longer appropriate, and we must reexamine the $z\leftrightarrow\zeta$ transformation. Integration of (2.1a) yields the relationship

$$\int_{\alpha^2}^{\zeta}\frac{(\xi-\alpha^2)^{1/2}}{2\xi}\,d\xi=\int_{x_t}^{z}\frac{\{t^2f(\alpha,t)\}^{1/2}}{t}\,dt. \tag{4.1}$$

It can be readily shown that the branches for the square roots and logarithmic singularities can be chosen so that $\zeta(z)$ is analytic at both $\zeta=0$ and $\zeta=\alpha^2$. We choose any branches satisfying this requirement.

The effect of the Liouville transformation is to yield the new form of differential equation

$$\frac{d^2W}{d\zeta^2}=\left\{u^2\left(\frac{1}{4\zeta}-\frac{\alpha^2}{4\zeta^2}\right)+\frac{\psi(\alpha,\zeta)}{\zeta}-\frac{1}{4\zeta^2}\right\}W, \tag{4.2}$$

where the new dependent variable is related to the original by

$$W(\zeta)=\left(\frac{\zeta-\alpha^2}{4\zeta^2f(\alpha,z)}\right)^{1/4}w(z). \tag{4.3}$$

Neglecting the term $\psi(\alpha,\zeta)/\zeta$ in (4.2) again results in the comparison equation (2.7) which has solutions $\zeta^{1/2}\mathscr{L}_{iu\alpha}(u\zeta^{1/2})$, where $\mathscr{L}$ again represents the modified Bessel functions I or K or any linear combination of the two. As in the real variable case we must select numerically satisfactory solutions. The major criterion is that all solutions that are recessive in the domain under consideration must be included.

Consider all modified Bessel functions $\mathscr{L}_{i\nu}(z)$ that are recessive when ν is positive and the complex argument[2] z takes its principal value $(-\pi<\arg(z)\leqq\pi)$. There are three such functions, namely, $K_{i\nu}(ze^{-\pi i})$, $K_{i\nu}(ze^{\pi i})$, and $K_{i\nu}(z)$, these being recessive at infinity for $\pi/2<\arg(z)\leqq\pi$, $-\pi<\arg(z)<\pi/2$, and $-\pi/2<\arg(z)<\pi/2$, respectively.

[2] The variable z here denotes a generic complex argument, and has no relation to that of (4.1).

We thus introduce at this stage the notation

$$\mathscr{L}_{i\nu}^{(j)}(z) = K_{i\nu}(z\,e^{-\pi i}),\ K_{i\nu}(z\,e^{\pi i}),\ 2\cosh(\nu\pi)K_{i\nu}(z), \tag{4.4}$$

for $j=1, 2, 3$, respectively, with the definition that $\mathscr{L}_{i\nu}^{(1)}(z)$ and $\mathscr{L}_{i\nu}^{(2)}(z)$ are to be continuous in the principal plane $-\pi < \arg(z) \leqq \pi$, taking their principal values respectively above and below the cut $\arg(z) = \pi$.

The factor $2\cosh(\nu\pi)$ in (4.4) was introduced for convenience; from the relations

$$K_{i\nu}(z\,e^{-\pi i}) = e^{-\nu\pi}K_{i\nu}(z) + \pi i I_{i\nu}(z), \qquad K_{i\nu}(z\,e^{\pi i}) = e^{\nu\pi}K_{i\nu}(z) - \pi i I_{i\nu}(z),$$

we find that the Wronskian

$$|\mathscr{W}\{\zeta^{1/2}\mathscr{L}_{i\nu}^{(j)}(u\zeta^{1/2}),\ \zeta^{1/2}\mathscr{L}_{i\nu}^{(j\pm1)}(u\zeta^{1/2})\}| = \pi\cosh(\nu\pi), \tag{4.5}$$

a result that will be used in the proof of the theorem on error bounds below. *Here and subsequently it will be supposed that j is enumerated* modulo 3.

When $\nu > 0$ no modified Bessel function $\mathscr{L}_{i\nu}(z)$ is recessive at $z=0$. When $\nu = 0$, however, the modified Bessel function $I_0(z)$ is uniquely characterized as being recessive at $z=0$; all other independent solutions, including $\mathscr{L}_0^{(j)}(z)$ $(j=1, 2, 3)$, are logarithmically singular at $z=0$. To complete our numerically satisfactory set we require, therefore, a fourth solution $\mathscr{L}_{i\nu}(z)$, which is proportional to $I_0(z)$ when $\nu = 0$. The function $\tilde{I}_{i\nu}(z)$ defined by (2.8) is not satisfactory, since it is not linearly independent of each of $\mathscr{L}_{i\nu}^{(j)}(z)$ for all values of ν in $0 \leqq \nu < \infty$. We find from the Wronskian of $\tilde{I}_{i\nu}(z)$ and $\mathscr{L}_{i\nu}^{(2)}(z)$ that these functions are multiples of each other when $\nu = (\ln 3)/(2\pi)$.

Our choice of a fourth numerically satisfactory solution then is simply the modified Bessel function $I_{i\nu}(z)$ itself, since it is linearly independent of each of the other three for all values of ν in $0 \leqq \nu < \infty$.

We shall seek formal series solutions of (4.2) in the form

$$\zeta^{1/2}\mathscr{L}_{iu\alpha}^{(j)}(u\zeta^{1/2}) \sum_{s=0}^{\infty} \frac{A_s(\alpha,\zeta)}{u^{2s}} + \frac{\zeta}{u}\mathscr{L}_{iu\alpha}^{(j)'}(u\zeta^{1/2}) \sum_{s=0}^{\infty} \frac{B_s(\alpha,\zeta)}{u^{2s}} \qquad (j=1,2,3), \tag{4.6a}$$

$$\zeta^{1/2}I_{iu\alpha}(u\zeta^{1/2}) \sum_{s=0}^{\infty} \frac{A_s(\alpha,\zeta)}{u^{2s}} + \frac{\zeta}{u}I'_{iu\alpha}(u\zeta^{1/2}) \sum_{s=0}^{\infty} \frac{B_s(\alpha,\zeta)}{u^{2s}}, \tag{4.6b}$$

where the coefficients $\{A_s\}$ and $\{B_s\}$ are the analytic continuations of those given by (2.13) and (2.14). The square root $\zeta^{1/2}$, which appears in (4.6a, b), is defined to satisfy

$$\Omega_0 < \arg\zeta^{1/2} \leqq \Omega_0 + \pi, \tag{4.7}$$

for some Ω_0 in $-\pi \leqq \Omega_0 \leqq 0$. The choice of Ω_0 depends on the particular application. We denote by $\underline{\Delta}$ the ζ-domain Δ having a cut along $\arg\zeta = 2\Omega_0$; we will seek solutions of the form (4.6a, b) in $\underline{\Delta}$, considering the series (4.6a) first.

In order to construct the desired error bounds we must define appropriate auxiliary functions for each of $\mathscr{L}_{i\nu}^{(j)}(z)$ $(j=1, 2, 3)$. In order to do this we must take into account the uniform asymptotic behavior of these functions in the z plane as $\nu \to \infty$; see [4, § 4]. With these results in mind we will shortly define weight functions. Before doing so we need to give a number of definitions. First define $\hat{X}_\nu \equiv X_{\hat{\nu}}$, where $\hat{\nu}$ denotes the real root of the equation

$$\nu = \hat{\nu} + c(\tfrac{1}{2}\hat{\nu})^{1/3}.$$

It is necessary to introduce this new turning point $\hat{X}_\nu$ in order that subsequent suprema appearing in error bounds should exist; see [1, Eq. (B.4)].

Note that $\hat{\nu} \leqq \nu$, and $\hat{\nu}$ increases monotonically from zero to infinity as ν increases from zero to infinity. Thus $\hat{X}_\nu \geqq \hat{X}_0 = X_0 > 0$. From (2.21) we observe that

$$\hat{X}_\nu = \nu + O(\nu^{-1/3}) \quad \text{as } \nu \to \infty. \tag{4.8}$$

Next define

$$\Phi_\nu^{(j)}(z) = \int_{\hat{X}_\nu}^{z} \frac{(t^2 - \hat{X}_\nu^2)^{1/2}}{t}\, dt \qquad (j = 1, 2, 3). \tag{4.9}$$

These functions have branchpoints at $z = 0, \pm\hat{X}_\nu$. With respect to the logarithmic branchpoint at $z = 0$ we introduce (for all three functions) a cut along $\arg z = \pi$; likewise, for the branchpoint at $z = -\hat{X}_\nu$ we introduce a cut along the real axis from $z = -\hat{X}_\nu$ to $z = -\infty$. For the branchpoint $z = \hat{X}_\nu$ we introduce three cuts. For $\Phi_\nu^{(0)}(z)$ we assign a cut along the real axis from $z = \hat{X}_\nu$ to $z = -\infty$.

Let Γ_1 denote the curve emanating from $z = \hat{X}_\nu$ in the first quadrant such that $\operatorname{Re} \Phi_\nu^{(j)}(z) = 0$. This is illustrated in Fig. 1; it makes an angle of $\pi/3$ with the real axis, and is asymptotic to the line $\operatorname{Re} z = (\pi/2)\hat{X}_\nu$ as $z \to \infty$. More precisely, we have parametrically as $z \to \infty$ on Γ_1

$$z(\tau) = i\tau + \frac{\pi}{2}\hat{X}_\nu - \frac{\pi \hat{X}_\nu^3}{4\tau^2} + O\left(\frac{1}{\tau^4}\right), \qquad \tau \to \infty. \tag{4.10}$$

Denote by Γ_2 the curve conjugate to Γ_1. We then define $\Phi_\nu^{(2)}(z)$ to have a cut along Γ_1, and $\Phi_\nu^{(1)}(z)$ to have a cut along Γ_2.

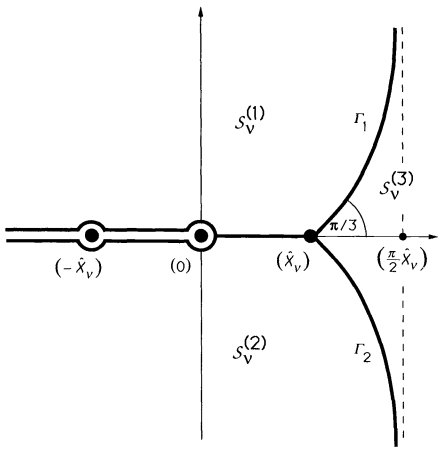


FIG. 1. *z plane.*

The cuts emanating from $\hat{X}_\nu$ and $-\hat{X}_\nu$ divide the z plane into three regions. We denote these domains (including their boundaries) by $\mathcal{S}_\nu^{(j)}$, $(j = 1, 2, 3)$ (see Fig. 1). The branches for $\Phi_\nu^{(j)}(z)$ are now chosen so that

$$\operatorname{Re} \Phi_\nu^{(j)}(z) > 0 \quad \text{for } z \in \mathcal{S}_\nu^{(j)}, \qquad \operatorname{Re} \Phi_\nu^{(j)}(z) < 0 \quad \text{for } z \in \mathcal{S}_\nu^{(j+1)} \cup \mathcal{S}_\nu^{(j-1)}.$$

We now are in a position to introduce the three weight functions. We define

$$E_\nu^{(j)}(z) = |\exp(\Phi_\nu^{(j)}(z))| \qquad (j = 1, 2, 3). \tag{4.11}$$

Thus, with our choice of branches for $\Phi_\nu^{(j)}(z)$, we see that $E_\nu^{(j)}(z) \geqq 1$ in $\mathcal{S}_\nu^{(j)}$ and $E_\nu^{(j)}(z) \leqq 1$ in $\mathcal{S}_\nu^{(j+1)} \cup \mathcal{S}_\nu^{(j-1)}$. This is as desired since, except in a neighborhood of $z = 0$ when $\nu = 0$, $\mathcal{L}_{i\nu}^{(j)}(z)$ is recessive in $\mathcal{S}_\nu^{(j)}$ and dominant in $\mathcal{S}_\nu^{(j+1)} \cup \mathcal{S}_\nu^{(j-1)}$.

Consider the behavior of the weight functions as $z \to 0$. We find from residue theory and (4.9) that the limiting behavior depends on the argument of z only, not on the modulus of z; setting $z = \varepsilon e^{i\phi} (-\pi < \phi < \pi)$ in (4.11) and letting $\varepsilon \to 0$ yield the following limits:

$$
\begin{aligned}
&E_\nu^{(1)}(\varepsilon e^{i\phi}) \to \exp\{\hat{X}_\nu \phi\}, \\
&E_\nu^{(2)}(\varepsilon e^{i\phi}) \to \exp\{-\hat{X}_\nu \phi\}, \\
&E_\nu^{(3)}(\varepsilon e^{i\phi}) \to \exp\{-\hat{X}_\nu |\phi|\}
\end{aligned} \tag{4.12}
$$

for each value of ν in $0 \leqq \nu < \infty$.

The level curves of $E_\nu^{(j)}(z)$ play an important role in the development and application of the following asymptotic expansions. From the definition of the weight functions we see that these are the family of curves defined by

$$
\operatorname{Re} \Phi_\nu^{(j)}(z) = k, \qquad -\infty < k < \infty, \tag{4.13}
$$

which include for $k = 0$ the curves Γ_1, Γ_2, and real interval $0 < z < \hat{X}_\nu$. The general configuration of these curves in the z plane is indicated in Fig. 2.

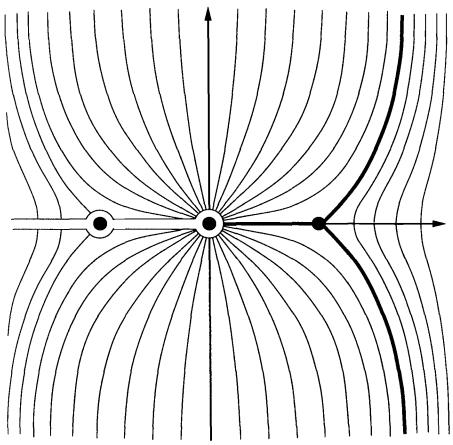


FIG. 2. *Level curves in z plane.*

The following observations can be noted for the level curves when $j = 1$. First, the curves in $\mathscr{S}_\nu^{(1)}$ have $k \geqq 0$, and in $\mathscr{S}_\nu^{(2)} \cup \mathscr{S}_\nu^{(3)}$ they have $k \leqq 0$. The value of k for each curve emanating from $z = 0$ is $\hat{X}_\nu \phi$, where ϕ is the (principal) angle the curve makes with the real z axis. The curves emanating from[3] $z = -\hat{X}_\nu \pm i0$ have $k = \pm \hat{X}_\nu \pi$. If $z = z_R$ denotes the intersection of a level curve with the real axis, then the curves in $\mathscr{S}_\nu^{(1)}$ are such that $k \to \infty$ as $|z_R| \to \infty$, and those in $\mathscr{S}_\nu^{(2)}$ or $\mathscr{S}_\nu^{(3)}$ are such that $k \to -\infty$ as $|z_R| \to \infty$. Finally, each curve is asymptotic at infinity to a line parallel to the imaginary axis. Similar arguments to those above apply when $j = 2$ and 3.

We next define modulus and phase functions by the relations

$$
E_\nu^{(j+1)}(z) |\mathscr{L}_{i\nu}^{(j+1)}(z)| = M_\nu^{(j)}(z) \sin\{\theta_\nu^{(j)}(z)\}, \tag{4.14a}
$$

$$
E_\nu^{(j-1)}(z) |\mathscr{L}_{i\nu}^{(j-1)}(z)| = M_\nu^{(j)}(z) \cos\{\theta_\nu^{(j)}(z)\}, \tag{4.14b}
$$

[3] The notation $\pm i0$ refers to points above and below a branchcut on the real axis.

where $M_\nu^{(j)}(z)$ is real and positive, and $\theta_\nu^{(j)}(z)$ is real. They are thus explicitly given by

$$M_\nu^{(j)}(z)=\{E_\nu^{(j+1)}(z)^2|\mathscr{L}_{i\nu}^{(j+1)}(z)|^2+E_\nu^{(j-1)}(z)^2|\mathscr{L}_{i\nu}^{(j-1)}(z)|^2\}^{1/2}, \tag{4.15}$$

$$\theta_\nu^{(j)}(z)=\arctan\left\{\frac{E_\nu^{(j+1)}(z)|\mathscr{L}_{i\nu}^{(j+1)}(z)|}{E_\nu^{(j-1)}(z)|\mathscr{L}_{i\nu}^{(j-1)}(z)|}\right\}, \tag{4.16}$$

where the inverse tangent takes its principal value. Note that for $\nu=0$ and $z\to 0$

$$M_0^{(j)}(z)\sim c_j(\phi)\ln|z|, \tag{4.17}$$

where

$$\begin{aligned} c_{1,2}(\phi)&=(\exp\{\mp 2\hat{X}_0\phi\}+4\exp\{-2\hat{X}_0|\phi|\})^{1/2},\\ c_3(\phi)&=(2\cosh\{2\hat{X}_0\phi\})^{1/2}, \end{aligned} \tag{4.18}$$

and $\phi\equiv\operatorname{Arg} z$.

We define modulus and phase functions for the derivatives by

$$E_\nu^{(j+1)}(z)|\mathscr{L}_{i\nu}^{(j+1)\prime}(z)|=N_\nu^{(j)}(z)\sin\{\omega_\nu^{(j)}(z)\}, \tag{4.19a}$$

$$E_\nu^{(j-1)}(z)|\mathscr{L}_{i\nu}^{(j-1)\prime}(z)|=N_\nu^{(j)}(z)\cos\{\omega_\nu^{(j)}(z)\}, \tag{4.19b}$$

where $N_\nu^{(j)}(z)$ is real and positive, and $\omega_\nu^{(j)}(z)$ is real.

The following constants will appear in the theorem on error bounds below. We define

$$\kappa^{(j\pm1)}=\sup\left\{\frac{\operatorname{sech}(\nu\pi)}{\pi}|z^2-\nu^2|^{1/2}M_\nu^{(j\pm1)}(z)^2\right\}, \tag{4.20}$$

$$\mu^{(j\pm1)}=\sup\left\{\frac{\operatorname{sech}(\nu\pi)}{\pi}|z^2-\nu^2|^{1/2}M_\nu^{(j\pm1)}(z)E_\nu^{(j\mp1)}(z)^{-1}|\mathscr{L}_\nu^{(j)}(z)|\right\}, \tag{4.21}$$

each supremum being evaluated over $\nu\geqq 0$ and $z\in\mathscr{S}_\nu^{(j)}\cup\mathscr{S}_\nu^{(j\mp1)}$. The existence of these suprema, and of those given by (4.32) and (4.33) below, can be established in a manner similar to the proof of Lemma 3 in [1, App. B].

Now let us focus our attention on the ζ plane. Under the transformation $z=u\zeta^{1/2}$, where the branch of the square root is given by (4.7) above, we define regions $S_\alpha^{(j)}$ in the ζ plane to be those corresponding to the regions $\mathscr{S}_{u\alpha}^{(j)}$ in the z plane. These regions are illustrated in Fig. 3a for $-\frac{1}{2}\pi<\Omega_0<0$, and it is seen that in this case $S_\alpha^{(2)}$ is bounded.

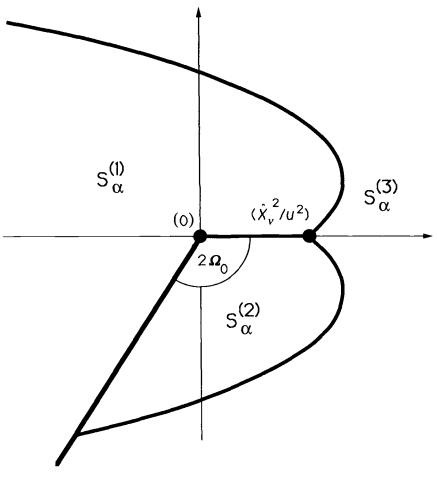


FIG. 3a. *ζ plane.*

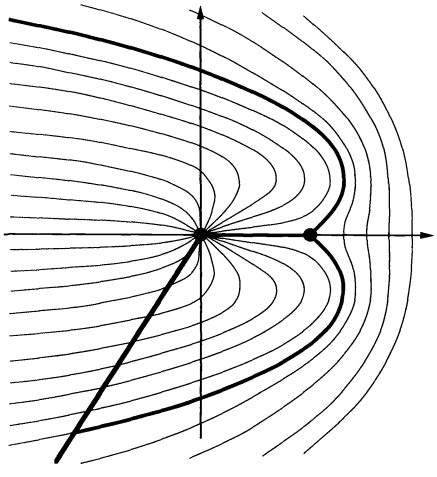


FIG. 3b. *Level curves in ζ plane.*

When $-\pi<\Omega_0<-\frac{1}{2}\pi$ the region $S_\alpha^{(3)}$ is bounded. Note also that $S_\alpha^{(2)}$ vanishes when $\Omega_0=0$, and $S_\alpha^{(3)}$ vanishes when $\Omega_0=-\pi$.

The error bounds below will be valid in subdomains $\underline{\Delta}^{(j)}$ of $\underline{\Delta}$, which are defined as follows. Let $\tilde{\zeta}^{(j)}\in S_\alpha^{(j)}$ $(j=1,2,3)$ denote three reference points chosen, possibly at infinity, to suit our purposes. The domains $\underline{\Delta}^{(j)}$ $(j=1,2,3)$ are defined to be the set of points in $\underline{\Delta}$ that can be linked to $\tilde{\zeta}^{(j)}$ by a path $\mathscr{P}^{(j)}$, which consists of a finite chain of R_2 arcs having the property that as ξ passes along $\mathscr{P}^{(j)}$ from $\tilde{\zeta}^{(j)}$ to ζ

$$\operatorname{Re}\Phi_{u\alpha}^{(j)}(u\xi^{1/2})\geqq\operatorname{Re}\Phi_{u\alpha}^{(j)}(u\zeta^{1/2}). \tag{4.22}$$

A sufficient condition for this to be true is for $\operatorname{Re}\Phi_{u\alpha}^{(j)}(u\xi^{1/2})$ to be nonincreasing as ξ passes from $\tilde{\zeta}^{(j)}$ to ζ. If this condition is met the path $\mathscr{P}^{(j)}$ is said to be progressive. A knowledge of the general configuration of the level curves $\operatorname{Re}\Phi_{u\alpha}^{(j)}(u\zeta^{1/2})=\text{constant}$ in the ζ plane is therefore helpful in determining the domains of validity $\underline{\Delta}^{(j)}$. These correspond to the level curves in the z plane (see Fig. 2), and are indicated in Fig. 3b. Each of the level curves in the ζ plane is asymptotic to a parabola at infinity.

We now are in a position to state our theorem on error bounds, which can be proved in a manner similar to Theorems 1 and 2.

THEOREM 3a. *With the conditions given in* § 1 *and the present section,* (4.2) *has, for each positive* u, *nonnegative* α, *and nonnegative integer* n, *the following three solutions* $(j=1,2,3)$, *which are holomorphic in* $\underline{\Delta}$ *and satisfy*

$$\begin{aligned}W_{2n+1}^{(j)}(u,\alpha,\zeta)&=\zeta^{1/2}\mathscr{L}_{iu\alpha}^{(j)}(u\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{u^{2s}}\\&\quad+\frac{\zeta}{u}\mathscr{L}_{iu\alpha}^{(j)'}(u\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{u^{2s}}+\varepsilon_{2n+1}^{(j)}(u,\alpha,\zeta),\end{aligned} \tag{4.23}$$

where

$$\begin{aligned}&\frac{|\varepsilon_{2n+1}^{(j)}(u,\alpha,\zeta)|}{\zeta^{1/2}M_{u\alpha}^{(j\pm1)}(u\zeta^{1/2})},\ \frac{|\partial\varepsilon_{2n+1}^{(j)}(u,\alpha,\zeta)/\partial\zeta|}{\frac{1}{2}uN_{u\alpha}^{(j\pm1)}(u\zeta^{1/2})+\frac{1}{2}\zeta^{-1/2}M_{u\alpha}^{(j\pm1)}(u\zeta^{1/2})}\\&\qquad\leqq\frac{\mu^{(j\pm1)}}{u^{2n+1}}E_{u\alpha}^{(j)}(u\zeta^{1/2})^{-1}\mathscr{V}_{\mathscr{P}^{(j)}}\{(\zeta-\alpha^2)^{1/2}B_n(\alpha,\zeta)\}\\&\qquad\quad\cdot\exp\left\{\frac{\kappa^{(j\pm1)}}{u}\mathscr{V}_{\mathscr{P}^{(j)}}\{(\zeta-\alpha^2)^{1/2}B_0(\alpha,\zeta)\}\right\},\end{aligned} \tag{4.24}$$

when $\zeta\in\underline{\Delta}^{(j)}$. *In* (4.24) *the suffix on* M, N, μ, *and* κ *is* $j+1$ *when* $\zeta\in S_\alpha^{(j)}\cup S_\alpha^{(j-1)}$, *and* $j-1$ *when* $\zeta\in S_\alpha^{(j)}\cup S_\alpha^{(j+1)}$.

When the domain $\underline{\Delta}$ is unbounded, Theorem 3a is uniformly valid for unbounded values of ζ, provided the variations of $(\zeta-\alpha^2)^{1/2}B_s(\alpha,\zeta)$ converge as $\zeta\to\infty$ in $\underline{\Delta}$.

It remains to construct an error bound corresponding to expansion (4.6b). To do this we again must introduce appropriate auxiliary functions. Bearing in mind that $I_{i\nu}(z)$ is dominant at infinity for $\nu\geqq0$, we assign for this function the weight function

$$\mathscr{E}_\nu(z)=|\exp(-\Phi_\nu^{(j)}(z))|,\qquad z\in\mathscr{S}_\nu^{(j)}\quad(j=1,2,3). \tag{4.25}$$

In constructing the error bound we shall consider the cases $z\in\mathscr{S}_\nu^{(j)}$ $(j=1,2,3)$ separately. In each domain we will use $\mathscr{L}_{i\nu}^{(j)}(z)$ as the companion solution to $I_{i\nu}(z)$. This time it is necessary to define new weight functions for $\mathscr{L}_{i\nu}^{(1)}(z)$ and $\mathscr{L}_{i\nu}^{(2)}(z)$ in order to take into account their logarithmic behavior at $z=0$ when $\nu=0$.

To this end we first define domains $\mathscr{D}_\nu^{(j)}(z)$ $(j=1,2)$ to be the set of points in the t plane satisfying the conditions

$$\operatorname{Re}\Phi_\nu^{(j)}(t) > \operatorname{Re}\Phi_\nu^{(j)}(z), \tag{4.26a}$$

$$|t| > \min\{|z|, \hat{X}_\nu\}, \tag{4.26b}$$

$$-\pi < \arg t < \pi. \tag{4.26c}$$

For example, the domain $\mathscr{D}_\nu^{(1)}(z)$ is the shaded region illustrated in Fig. 4 (for the case $|z| > \hat{X}_\nu$).

We now define for $j=1, 2$

$$\mathscr{E}_\nu^{(j)}(z)^{-1} = \frac{1}{a_\nu^{(j)}} \sup_{t\in\mathscr{D}_\nu^{(j)}(z)} \{|t^2-\nu^2|^{1/4}|\mathscr{L}_{i\nu}^{(j)}(t)|\}, \qquad z\in\mathscr{S}_\nu^{(j)}), \tag{4.27}$$

where the normalizing coefficient $a_\nu^{(j)}$ is chosen so that $\mathscr{E}_\nu^{(j)}(z)=1$ for $z\in\Gamma_j$. Thus

$$a_\nu^{(j)} = \sup_{t\in\mathscr{D}_\nu^{(j)}(\chi_\nu)} \{|t^2-\nu^2|^{1/4}|\mathscr{L}_{i\nu}^{(j)}(t)|\}. \tag{4.28}$$

Since for all $\nu \geqq 0$ the singularity $\{0\} \notin \mathscr{S}_\nu^{(3)}$, it is not necessary to introduce a different weight function for $\mathscr{L}_{i\nu}^{(3)}(z)$. Thus

$$\mathscr{E}_\nu^{(3)}(z) = E_\nu^{(3)}(z). \tag{4.29}$$

Note that, for $j=1$ and 2, it has been necessary to define only $\mathscr{E}_\nu^{(j)}(z)$ for $z\in\mathscr{S}_\nu^{(j)}$.

Real and positive modulus functions $\mathscr{M}_\nu^{(j)}(z)$ and real phase functions $\Theta_\nu^{(j)}(z)$ are now defined for $z\in\mathscr{S}_\nu^{(j)}$ by

$$\mathscr{E}_\nu(z)|I_{i\nu}(z)| = \mathscr{M}_\nu^{(j)}(z)\sin\{\Theta_\nu^{(j)}(z)\}, \tag{4.30a}$$

$$\mathscr{E}_\nu^{(j)}(z)|\mathscr{L}_{i\nu}^{(j)}(z)| = \mathscr{M}_\nu^{(j)}(z)\cos\{\Theta_\nu^{(j)}(z)\}. \tag{4.30b}$$

Likewise, for the derivatives we define real and positive modulus functions $\mathscr{N}_\nu^{(j)}(z)$, and real phase functions $\bar{\omega}_\nu^{(j)}(z)$, by

$$\mathscr{E}_\nu(z)|I'_{i\nu}(z)| = \mathscr{N}_\nu^{(j)}(z)\sin\{\bar{\omega}_\nu^{(j)}(z)\}, \tag{4.31a}$$

$$\mathscr{E}_\nu^{(j)}(z)|\mathscr{L}_{i\nu}^{(j)'}(z)| = \mathscr{N}_\nu^{(j)}(z)\cos\{\bar{\omega}_\nu^{(j)}(z)\}. \tag{4.31b}$$

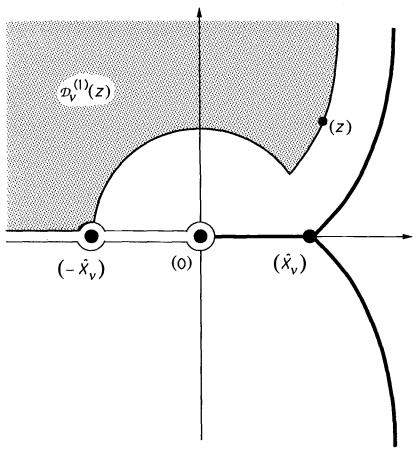


FIG. 4. *t plane.*

For correspondence to (4.20) and (4.21) we introduce the following constants:

$$\kappa_0^{(j)} = \sup\{\lambda^{(j)}(\nu)|z^2-\nu^2|^{1/2}\mathscr{M}_\nu^{(j)}(z)^2\mathscr{E}_\nu(z)^{-1}\mathscr{E}_\nu^{(j)}(z)^{-1}\}, \tag{4.32}$$

$$\mu_0^{(j)} = \sup\{\lambda^{(j)}(\nu)|z^2-\nu^2|^{1/2}\mathscr{M}_\nu^{(j)}(z)\mathscr{E}_\nu^{(j)}(z)^{-1}|I_{i\nu}(z)|\}, \tag{4.33}$$

each supremum being evaluated over $\nu \geqq 0$ and $z \in \mathscr{S}_\nu^{(j)}$. In these equations $\lambda^{(j)}(\nu)$ are constants involving Wronskians, viz.

$$\lambda^{(j)}(\nu) = |\mathscr{W}\{\zeta^{1/2}\mathscr{L}_{i\nu}^{(j)}(u\zeta^{1/2}), \zeta^{1/2}I_{i\nu}(u\zeta^{1/2})\}|^{-1}.$$

Thus

$$\lambda^{(1)}(\nu) = 2\,e^{\nu\pi}, \quad \lambda^{(2)}(\nu) = 2\,e^{-\nu\pi}, \quad \lambda^{(3)}(\nu) = \operatorname{sech}(\nu\pi). \tag{4.34}$$

Next we define, for $j=1$ and 2, a domain $\underline{\Delta}_0^{(j)}$ as being the set of points in $S_\alpha^{(j)}$ that can be linked to $\zeta=0$ by a certain path $\mathscr{P}_0^{(j)}$ in $\underline{\Delta}$. The path $\mathscr{P}_0^{(j)}$ is one consisting of a finite chain of R_2 arcs, and has the properties that as ξ passes along $\mathscr{P}_0^{(j)}$ from zero to $\zeta(\in S_\alpha^{(j)})$

$$\operatorname{Re}\Phi_{u\alpha}^{(j)}(u\xi^{1/2}) \leqq \operatorname{Re}\Phi_{u\alpha}^{(j)}(u\zeta^{1/2}), \tag{4.35}$$

$$|u\xi^{1/2}| \leqq |u\zeta^{1/2}|. \tag{4.36}$$

Similarly, we define $\underline{\Delta}_0^{(3)}$ as the set of points in $S_\alpha^{(3)}$ that can be linked to $\zeta=0$ by a path $\mathscr{P}_0^{(3)}$ in $\underline{\Delta}$ consisting of a finite chain of R_2 arcs, and has the property that as ξ passes along $\mathscr{P}_0^{(3)}$ from zero to $\zeta(\in S_\alpha^{(3)})$

$$\operatorname{Re}\Phi_{u\alpha}^{(3)}(u\xi^{1/2}) \leqq \operatorname{Re}\Phi_{u\alpha}^{(3)}(u\zeta^{1/2}). \tag{4.37}$$

If all points in the interval $[0, \hat{X}_\nu^2/u^2]$ are in $\underline{\Delta}$, then a natural choice for $\mathscr{P}_0^{(3)}$ would be the path consisting of the real interval $[0, \hat{X}_\nu^2/u^2]$ linked to a progressive path in $S_\alpha^{(3)}$.

We now state the following theorem on error bounds for a solution that is recessive at $\zeta=0$ when $\alpha=0$.

THEOREM 3b. *With the conditions given in* § 1 *and the present section,* (4.2) *has, for each positive u, nonnegative* α*, and nonnegative integer n, a solution that is holomorphic in* $\underline{\Delta}$ *satisfying*

$$\begin{aligned} W_{2n+1}(u,\alpha,\zeta) &= \zeta^{1/2}I_{iu\alpha}(u\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{u^{2s}} \\ &\quad + \frac{\zeta}{u}I'_{iu\alpha}(u\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{u^{2s}} + \varepsilon_{2n+1}(u,\alpha,\zeta), \end{aligned} \tag{4.38}$$

where the following bounds hold for $\zeta \in \underline{\Delta}_0^{(j)}$,

$$\begin{aligned} &\frac{|\varepsilon_{2n+1}(u,\alpha,\zeta)|}{\zeta^{1/2}\mathscr{M}_{u\alpha}^{(j)}(u\zeta^{1/2})}, \frac{|\partial\varepsilon_{2n+1}(u,\alpha,\zeta)/\partial\zeta|}{\frac{1}{2}u\mathscr{N}_{u\alpha}^{(j)}(u\zeta^{1/2})+\frac{1}{2}\zeta^{-1/2}\mathscr{M}_{u\alpha}^{(j)}(u\zeta^{1/2})} \\ &\leqq \frac{\mu_0^{(j)}}{u^{2n+1}}\mathscr{E}_\nu(u\zeta^{1/2})^{-1}\mathscr{V}_{\mathscr{P}_0^{(j)}}\{(\zeta-\alpha^2)^{1/2}B_n(\alpha,\zeta)\} \\ &\quad \cdot \exp\left\{\frac{\kappa_0^{(j)}}{u}\mathscr{V}_{\mathscr{P}_0^{(j)}}\{(\zeta-\alpha^2)^{1/2}B_0(\alpha,\zeta)\}\right\}. \end{aligned} \tag{4.39}$$

As before, Theorem 3b is valid for unbounded ζ, provided the variations of $(\zeta-\alpha^2)^{1/2}B_s(\alpha,\zeta)$ converge as $\zeta\to\infty$.

Finally, we remark that discussions similar to those following Theorems 1 and 2 hold for the asymptotic behavior of the expansions (4.6a, b); see also [1, eqs. (5.17)-(5.19)].

5. Legendre functions of purely imaginary order and real argument and degree. As an illustration of the preceding theory we shall derive asymptotic expansions, as $\nu \to \infty$, of solutions of the following form of the associated Legendre equation

$$\frac{d^2w}{dx^2} = \left\{ -\frac{\nu(\nu+1)}{1-x^2} - \frac{\mu^2+1}{(1-x^2)^2} \right\} w. \tag{5.1}$$

In this section we shall suppose that the variables x, ν, μ are real, and hence so too is the differential equation (5.1). We restrict our attention to the case where $x \geqq 0$, $\nu > -\frac{1}{2}$, and $\mu \geqq 0$. Corresponding results for other real values of these variables (except $\nu = -\frac{1}{2}$) can be obtained using appropriate connection formulas. In § 6 we tackle the case where the argument x is complex.

We consider the intervals $0 \leqq x < 1$ and $1 < x < \infty$ separately. Solutions of (5.1) in the former interval are the Ferrers functions $(1-x^2)^{1/2}\mathrm{P}_\nu^{i\mu}(x)$, $(1-x^2)^{1/2}\mathrm{Q}_\nu^{i\mu}(x)$, or any linear combination of them. For $x > 1$ solutions are linear combinations of the Legendre functions $(x^2-1)^{1/2}P_\nu^{i\mu}(x)$ and $(x^2-1)^{1/2}\mathbf{Q}_\nu^{i\mu}(x)$. For definitions of these standard functions see [6, pp. 170, 185].

Equation (5.1) can be identified with (1.1) by defining

$$u = \nu + \frac{1}{2}, \qquad \alpha = \frac{\mu}{\nu+\frac{1}{2}}, \tag{5.2}$$

$$f(\alpha, x) = -\frac{\alpha^2}{(1-x^2)^2} - \frac{1}{1-x^2}, \qquad g(\alpha, x) = -\frac{1}{(1-x^2)^2} + \frac{1}{4(1-x^2)}. \tag{5.3}$$

Thus for $x \geqq 0$ the equation has regular singularities at 1 and infinity, and a turning point at $x_t = (1+\alpha^2)^{1/2}$. It is necessary for us to impose the restriction $0 \leqq \alpha \leqq A$, i.e.,

$$0 \leqq \frac{\mu}{\nu+\frac{1}{2}} \leqq A, \tag{5.4}$$

where A is an arbitrary positive constant. Thus we consider the case where the turning point x_t is bounded, and in the limiting case $\alpha \to 0$ coalesces with the singularity at $x = 1$.

We now apply the Liouville transformation described in § 2. From (2.2a, b) we define a new independent variable ζ by

$$\int_{\alpha^2}^{\zeta} \frac{(\xi-\alpha^2)^{1/2}}{2\xi}\, d\xi = \int_{(1+\alpha^2)^{1/2}}^{x} \frac{(t^2-1-\alpha^2)^{1/2}}{t^2-1}\, dt \qquad (x > (1+\alpha^2)^{1/2} \text{ or } \zeta > \alpha^2), \tag{5.5a}$$

$$\int_{\alpha^2}^{\zeta} \frac{(\alpha^2-\xi)^{1/2}}{2\xi}\, d\xi = \int_{(1+\alpha^2)^{1/2}}^{x} \frac{(1+\alpha^2-t^2)^{1/2}}{t^2-1}\, dt \qquad (x < (1+\alpha^2)^{1/2} \text{ or } \zeta < \alpha^2), \tag{5.5b}$$

where the Cauchy principal values are taken for (5.5b). Note that the singularities $x = 1$ and $x = \infty$ are mapped to $\zeta = 0$ and $\zeta = \infty$, respectively, and the turning point $x = (1+\alpha^2)^{1/2}$ is mapped to $\zeta = \alpha^2$.

The ζ integrals are given explicitly by (2.3a, b). The x integrals can also be evaluated, and their explicit forms will be of use below. We obtain for the right-hand sides (RHS) of (5.5a, b), respectively,

$$\ln\{(x^2-1-\alpha^2)^{1/2}+x\} - \frac{1}{2}\ln\{1+\alpha^2\} - \alpha \arctan\left\{\frac{(x^2-1-\alpha^2)^{1/2}}{\alpha x}\right\}, \tag{5.6a}$$

$$\frac{\alpha}{2}\ln\left\{\frac{x+1+\alpha^2+\alpha(1+\alpha^2-x^2)^{1/2}}{-x+1+\alpha^2+\alpha(1+\alpha^2-x^2)^{1/2}}\right\} + \frac{\alpha}{2}\ln\left\{\frac{|x-1|}{x+1}\right\} + \arccos\left\{\frac{x}{(1+\alpha^2)^{1/2}}\right\}. \tag{5.6b}$$

Here the inverse trigonometric functions take their principal values.

From (2.2b) and (5.6b) we see that the x interval $[0, \infty)$ is mapped to the ζ interval $[\zeta_1, \infty)$, where ζ_1 is the negative root of the equation

$$(\alpha^2-\zeta_1)^{1/2}-\frac{\alpha}{2}\ln\left\{\frac{\alpha+(\alpha^2-\zeta_1)^{1/2}}{(\alpha^2-\zeta_1)^{1/2}-\alpha}\right\}=\frac{\pi}{2}.$$

We see that the $x-\zeta$ transformation is quite complicated (as compared, say, to the corresponding transformations involved in a Liouville-Green approximation), and this may be regarded as the price we pay for the uniformity of the resulting asymptotic solutions.

To complete the Liouville transformation for this problem we define the new dependent variable W by

$$w=\left(\frac{\zeta-\alpha^2}{x^2-1-\alpha^2}\right)^{1/4}\left(\frac{x^2-1}{2\zeta}\right)^{1/2}W, \tag{5.7}$$

and as a result (5.1) is transformed to the desired form (1.2). The Schwarzian ψ is readily found to be given by

$$\psi(\alpha,\zeta)=\frac{\zeta+4\alpha^2}{16(\zeta-\alpha^2)^2}-\frac{(x^2-1)(\zeta-\alpha^2)\{(4\alpha^2+1)x^2+\alpha^4-1\}}{16\zeta(x^2-1-\alpha^2)^3}. \tag{5.8}$$

In the identification of the asymptotic solutions it is necessary to know the behavior of the $x-\zeta$ transformation near $x=1$ and $x=\infty$. First, from (2.2b) and (5.5b) we find $\zeta\to 0$ as $x\to 1$ such that

$$x-1=\frac{\exp\{T(\alpha)\}}{2(1+\alpha^2)}\zeta+O(\zeta^2), \tag{5.9}$$

where

$$T(\alpha)=2-\frac{2}{\alpha}\arctan(\alpha). \tag{5.10}$$

This result is uniformly valid for $0\leqq\alpha\leqq A$. In particular, when $\alpha=0$ the limiting form $x-1\sim\frac{1}{2}\zeta$ applies.

Next, from (2.2a) and (5.5a) we find $\zeta\to\infty$ as $x\to\infty$ such that

$$x=\tfrac{1}{2}(1+\alpha^2)^{1/2}\exp\{\zeta^{1/2}-V(\alpha)\}(1+O(\zeta^{-1/2})), \tag{5.11}$$

where

$$V(\alpha)=\alpha\arctan(\alpha). \tag{5.12}$$

Again when $\alpha=0$ the limiting form $x\sim\frac{1}{2}\exp\{\zeta^{1/2}\}$ applies.

An application of Theorem 2 to the transformed equation (1.2) yields the two real solutions $W_{2n+1,3}$ and $W_{2n+1,4}$. Let us identify the solution $W_{2n+1,3}$ with Ferrers functions. For these functions it is convenient to set the integration constants that appear in (2.14) so that

$$A_0(\alpha,\zeta)\equiv 1,\quad A_s(\alpha,0)=0\ \ (s=1,2,\cdots). \tag{5.13}$$

To facilitate identification we introduce two new parameters, defined by

$$\begin{aligned}\omega&\equiv\frac{u\alpha}{2}\ln\{u^2(1+\alpha^2)\}-\frac{u\alpha}{2}T(\alpha)\\&=\frac{\mu}{2}\ln\{(\nu+\tfrac{1}{2})^2+\mu^2\}-\mu+u\arctan\{\mu/u\},\end{aligned} \tag{5.14a}$$

$$\beta_n\equiv\arctan\left\{\frac{2\mu}{u^2}\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{u^{2s}}\right\}. \tag{5.14b}$$

With these definitions we introduce the following Ferrers function:

$$\mathbf{P}_n(\nu, \mu, x) \equiv \tfrac{1}{2}\{e^{i(\omega+\beta_n)}\mathrm{P}_\nu^{-i\mu}(x) + e^{-i(\omega+\beta_n)}\mathrm{P}_\nu^{i\mu}(x)\}, \tag{5.15}$$

which, unlike $\mathrm{P}_\nu^{i\mu}(x)$, is a real solution. Note also that

$$\mathbf{P}_n(\nu, 0, x) = \mathrm{P}_\nu(x), \tag{5.16}$$

and hence $\mathbf{P}_n(\nu, \mu, x)$ is recessive at $x=1$ when $\mu = 0$.

From (5.15) and the known behavior of Ferrers functions near $x=1$ (see, e.g., [6, p. 186]), it is seen that as $x \to 1^-$ $(\mu > 0)$

$$\mathbf{P}_n(\nu, \mu, x) \sim \left(\frac{\sinh(\mu\pi)}{\mu\pi}\right)^{1/2} \cos\left\{\frac{\mu}{2}\ln\left(\frac{1-x}{2}\right) - \phi_{\mu,0} + \omega + \beta_n\right\}, \tag{5.17}$$

where

$$\phi_{\mu,0} = \arg\Gamma(1+i\mu). \tag{5.18}$$

The phase $\phi_{\mu,0}$ is defined to be continuous in μ, such that $\phi_{0,0} = 0$. The limiting form of (5.17) applies when $\mu = 0$. From (5.2) and (5.9) we find that in terms of the new variables

$$\mathbf{P}_n(\nu, \mu, x) \sim \left(\frac{\sinh(u\alpha\pi)}{u\alpha\pi}\right)^{1/2} \cos\{u\alpha \ln(u|\zeta|^{1/2}) - \phi_{u\alpha,0} + \beta_n\}, \tag{5.19}$$

as $\zeta \to 0^-$.

Consider now the asymptotic solution of (1.2):

$$\left(\frac{\alpha^2-\zeta}{1+\alpha^2-x^2}\right)^{1/4} |\zeta|^{-1/2} W_{2n+1,3}(u, \alpha, \zeta). \tag{5.20}$$

The behavior of this function near $\zeta = 0$ can be established from (3.15) and the corresponding behavior of $F_{i\mu}(u|\zeta|^{1/2})$ (see [4, eq. (3.15)]). We find as $\zeta \to 0^-$ that (5.20) is of the form

$$\left(\frac{2\tanh(\mu\pi/2)}{\mu\pi}\right)^{1/2}\Bigg[\cos\{u\alpha \ln(u|\zeta|^{1/2}) - \phi_{u\alpha,0}\} - \frac{2\mu}{u^2}\sin\{u\alpha \ln(u|\zeta|^{1/2}) - \phi_{u\alpha,0}\}\sum_{s=0}^{n-1}\frac{B_s(\alpha, 0)}{u^{2s}} + O(\zeta)\Bigg],$$

or equivalently

$$\left(\frac{2\tanh(\mu\pi/2)}{\mu\pi}\right)^{1/2}\left(1+\left(\frac{2\mu}{u^2}\sum_{s=0}^{n-1}\frac{B_s(\alpha, 0)}{u^{2s}}\right)^2\right)^{1/2} \cos\{u\alpha \ln(u|\zeta|^{1/2}) - \phi_{u\alpha,0} + \beta_n\} + O(\zeta), \tag{5.21}$$

where β_n is defined by (5.14b). On comparing (5.21) with (5.19) we conclude that the two solutions must be identical to within a multiplicative constant. Therefore

$$\left(\frac{\alpha^2-\zeta}{1+\alpha^2-x^2}\right)^{1/4} |\zeta|^{-1/2} W_{2n+1,3}(u, \alpha, \zeta) = R_{2n+1,3}(\nu, \mu)\mathbf{P}_n(\nu, \mu, x), \tag{5.22}$$

where $R_{2n+1,3}$ is independent of x. If we compare both sides at $\zeta = 0$ we find that

$$R_{2n+1,3}(\nu, \mu) = \frac{1}{\cosh(\mu\pi/2)}\left(1+\left(\frac{2\mu}{u^2}\sum_{s=0}^{n-1}\frac{B_s(\alpha, 0)}{u^{2s}}\right)^2\right)^{1/2}. \tag{5.23}$$

As a useful independent check on the correctness of this asymptotic result we can show that at $x=0$ the asymptotic behavior as $u\to\infty$ of both sides of (5.22) is identical. This can be done as in the corresponding case in [1, § 4], using the well-known formula for Ferrers functions at the origin (see, e.g., [6, p. 187]), and the asymptotic formula for $F_{i\mu}(u|\zeta|^{1/2})$ having large order and argument (see [4, eq. (5.15)]).

Provided $\alpha>0$, the solution $W_{2n+1,4}$ is uniquely characterized by its oscillatory behavior at $\zeta=0$, and therefore it can also be identified with Ferrers functions. The coefficients in the resulting equation involve unknown terms that have explicit upper bounds, and are $O(u^{-2n-1})$ as $u\to\infty$. However, as $\alpha\to 0$ it turns out that one of the bounds no longer holds, and as such the asymptotic expansion is not uniformly valid in $0\leqq\alpha\leqq A$. This is perhaps as we would expect, since $W_{2n+1,4}$ is dominant at $\zeta=0$ when $\alpha=0$ and is therefore not uniquely characterized. For this reason we will defer the identification of a second independent solution in $0\leqq x<1$ until § 6, using complex variable results to do so.

For correspondence to (5.15) we define for $x>1$

$$P_n(\nu,\mu,x)\equiv\tfrac{1}{2}\{e^{i(\omega+\beta_n)}P_\nu^{-i\mu}(x)+e^{-i(\omega+\beta_n)}P_\nu^{i\mu}(x)\}. \tag{5.24}$$

This Legendre function is real, and can be shown in a manner similar to the example above to have the following asymptotic expansion:

$$\left(\frac{\zeta-\alpha^2}{x^2-1-\alpha^2}\right)^{1/4}\zeta^{-1/2}W_{2n+1,1}(u,\alpha,\zeta)=R_{2n+1,1}(\nu,\mu)P_n(\nu,\mu,x), \tag{5.25}$$

where

$$R_{2n+1,1}(\nu,\mu)=2\pi\,e^{-\mu\pi}\left(1+\left(\frac{2\mu}{u^2}\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{u^{2s}}\right)^2\right)^{1/2}. \tag{5.26}$$

The integration constants associated with the coefficients $\{A_s\}$ in $W_{2n+1,1}$ are the same as those for the function $W_{2n+1,3}$. Note that (5.8) and (5.11) imply that the Schwarzian ψ, given by (5.8), has the property that its sth derivative $\psi^{(s)}(\alpha,\zeta)=O(\zeta^{-s-1})$ as $\zeta\to\infty$. Thus the asymptotic solutions $W_{2n+1,1}$ and $W_{2n+1,2}$ given by Theorem 2 are uniformly valid for $0<\zeta<\infty$.

Finally, we identify the asymptotic solution given by (2.35) of Theorem 1. There exists a constant $R_{2n+1,2}$ such that

$$\left(\frac{\zeta-\alpha^2}{x^2-1-\alpha^2}\right)^{1/4}\zeta^{-1/2}W_{2n+1,2}(u,\alpha,\zeta)=R_{2n+1,2}(\nu,\mu)\mathbf{Q}_\nu^{i\mu}(x), \tag{5.27}$$

since the functions on both sides satisfy the same differential equation and have the unique property of being recessive as $x\to\infty$ ($\zeta\to\infty$). We choose the integration constants in (2.14) so that

$$A_s(\infty)=0\qquad(s=1,2,\cdots), \tag{5.28}$$

and determine the constant of proportionality $R_{2n+1,2}$ by comparing both sides of (5.27) as $\zeta\to\infty$. From (2.35), (5.2), (5.11), (5.12), and [4, eq. (2.16)] we find by this method

$$R_{2n+1,2}=\frac{\Gamma(\nu+\frac{3}{2})((\nu+\frac{1}{2})^2+\mu^2)^{\nu/2+1/4}\,e^{-(\nu+1/2)V}}{(\nu+\frac{1}{2})^{\nu+1}}\left[1-\frac{1}{\nu+\frac{1}{2}}\sum_{s=0}^{n-1}\frac{\mathscr{B}_s}{(\nu+\frac{1}{2})^{2s}}\right], \tag{5.29}$$

where

$$\mathscr{B}_s=\lim_{\zeta\to\infty}\zeta^{1/2}B_s(\alpha,\zeta). \tag{5.30}$$

As a check on the results in $\zeta>0$ we can compare both sides of (5.25) at $\zeta=\infty$, and both sides of (5.27) at $\zeta=0$, as $u\to\infty$. First, on employing (5.24) and the asymptotic

formula for $P_\nu^{\pm i\mu}(x)$ at $x=\infty$ (see, for example, [6, p. 173]) we find that the right-hand side (RHS) of (5.25) is of the form

$$\left(\frac{2\pi}{ux}\right)^{1/2}\frac{(2x)^u\exp\{u\alpha\arctan(\alpha)-u\alpha\pi\}}{(1+\alpha^2)^{u/2}}(1+o(1)) \tag{5.31}$$

for fixed positive α and large positive u and x. With the aid of (5.11) we find that, to leading order, the corresponding asymptotic behavior of the left-hand side (LHS) of (5.25) is identical.

On using (5.9), (5.24), and [6, p. 171, eqs. (12.08), (12.11)] the RHS of (5.27) is found to be of the form

$$-\left(\frac{2\pi}{u\alpha}\right)^{1/2}\exp\{-u\alpha\pi/2\}\sin\{u\alpha\ln(\zeta)/2-u\alpha\ln(2\alpha)+u\alpha-\pi/4\}(1+o(1)), \tag{5.32}$$

for fixed positive α, small ζ, and large positive u. Again, this is found to be in agreement with the corresponding behavior of the LHS of (5.27) (see [4, eqs. (2.14), (2.29)]).

The derivation of both (5.31) and (5.32) involved a considerable amount of algebra, which was eased somewhat by taking logarithms at appropriate stages and later exponentiating.

6. Legendre functions of purely imaginary order, real degree, and complex argument. In this section we will construct asymptotic expansions as $\nu\to\infty$ for the Legendre functions of complex argument $P_\nu^{-i\mu}(z)$ and $\mathbf{Q}_\nu^{i\mu}(z)$. The results will be uniformly valid for the parameters ν and μ satisfying (5.4), and for simplicity we assume $\operatorname{Re}(z)\geqq 0$. Corresponding results for other values of z may be readily obtained using the appropriate connection formulas for Legendre functions.

With u and α defined as before, the Liouville transformation in this case is given by

$$\int_{\alpha^2}^{\zeta}\frac{(\xi-\alpha^2)^{1/2}}{2\xi}\,d\xi=\int_{(1+\alpha^2)^{1/2}}^{z}\frac{(t^2-1-\alpha^2)^{1/2}}{t^2-1}\,dt. \tag{6.1}$$

The effect of this transformation is to map the half plane $\operatorname{Re}(z)\geqq 0$ to the domain Δ illustrated in Fig. 5a. The branchcut associated with the singularities at $z=\pm 1$ for

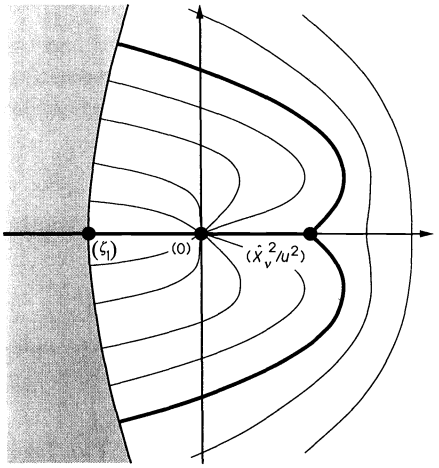


FIG. 5a. *Domain Δ in ζ plane.*

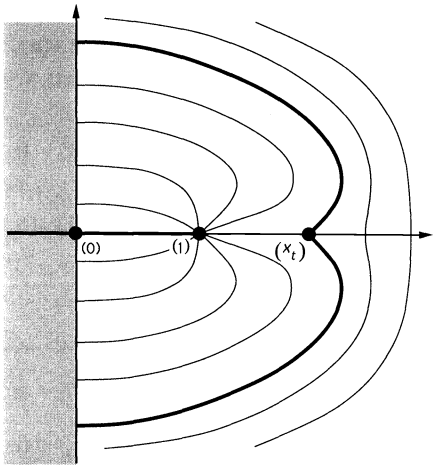


FIG. 5b. *z plane.*

Legendre functions is customarily taken to run along the real axis from $z=1$ to $z=-\infty$. Thus in the half plane $\mathrm{Re}\,(z)\geqq 0$ we take the cut associated with $z=1$ to be the interval $[0,1]$. The corresponding cut in the domain $\underline{\Delta}$ runs from $\zeta=0$ to $\zeta=\zeta_1$. In the notation of § 4 we thus have $\Omega_0=-\pi/2$.

The curves in the z plane corresponding to the level curves in the ζ plane are indicated in Fig. 5b. From the configuration of the level curves in the ζ plane it is readily verified that $\underline{\Delta}_0^{(1)}\cup\underline{\Delta}_0^{(2)}\cup\underline{\Delta}_0^{(3)}=\underline{\Delta}$. Also, on taking the reference point for $W_{2n+1}^{(3)}$ to be $\tilde{\zeta}^{(3)}=\infty$ we find that $\underline{\Delta}^{(3)}=\underline{\Delta}$. The asymptotic solutions W_{2n+1} and $W_{2n+1}^{(3)}$ are therefore uniformly valid throughout $\underline{\Delta}$.

The identification of the first of these follows in a similar manner to that of (5.25). From the behavior of the following solutions at $z=1$ we deduce that

$$\left(\frac{\zeta-\alpha^2}{z^2-1-\alpha^2}\right)^{1/4}\zeta^{-1/2}W_{2n+1}(u,\alpha,\zeta)=R_{2n+1}(\nu,\mu)P_\nu^{-i\mu}(z), \tag{6.2}$$

where, if we assign the integration constants as in (5.13), the proportionality constant is given by

$$R_{2n+1}(\nu,\mu)=e^{-i\mu T/2}((\nu+\tfrac{1}{2})^2+\mu^2)^{i\mu/2}\left[1+\frac{4i\mu}{2\nu+1}\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{(\nu+\frac{1}{2})^{2s}}\right]. \tag{6.3}$$

Likewise, we can show that the following relationship holds:

$$\left(\frac{\zeta-\alpha^2}{z^2-1-\alpha^2}\right)^{1/4}\zeta^{-1/2}W_{2n+1}^{(3)}(u,\alpha,\zeta)=R_{2n+1}^{(3)}(\nu,\mu)\mathbf{Q}_\nu^{i\mu}(z), \tag{6.4}$$

since both solutions are recessive at $z=\infty$ ($\zeta=\infty$). By comparing both sides at infinity, and assigning the integration constants so that (5.28) holds, we find that

$$R_{2n+1}^{(3)}(\nu,\mu)=\frac{2\cosh(\mu\pi)((\nu+\frac{1}{2})^2+\mu^2)^{(2\nu+1)/4}\Gamma(\nu+\frac{3}{2})\,e^{-(\nu+1/2)V}}{(\nu+\frac{1}{2})^{\nu+1}}\cdot\left[1-\frac{1}{\nu+\frac{1}{2}}\sum_{s=0}^{n-1}\frac{\mathscr{B}_s}{(\nu+\frac{1}{2})^{2s}}\right], \tag{6.5}$$

where the coefficients $\{\mathscr{B}_s\}$ are defined by (5.30).

It remains to tackle the problem left unsolved in § 5, namely, the identification of a second real asymptotic solution in the interval $[\zeta_1,0)$. It is convenient to define the following real Ferrers function:

$$\tilde{\mathbf{Q}}_\nu^{i\mu}(x)\equiv-\frac{1}{\pi}\{\mathbf{Q}_\nu^{i\mu}(x+i0)+\mathbf{Q}_\nu^{i\mu}(x-i0)\}, \tag{6.6}$$

and derive an asymptotic expansion for this function. Note that in terms of standard Ferrers functions

$$\tilde{\mathbf{Q}}_\nu^{i\mu}(x)=-\frac{1}{\pi\cosh(\mu\pi/2)}\left\{\frac{1}{\Gamma(\nu+i\mu+1)}\mathbf{Q}_\nu^{i\mu}(x)+\frac{1}{\Gamma(\nu-i\mu+1)}\mathbf{Q}_\nu^{-i\mu}(x)\right\}; \tag{6.7}$$

see, for example, [6, p. 185, eq. (15.02)]. Employing (6.4)-(6.6) and [4, eq. (3.9)] we derive the following asymptotic expansion:

$$\begin{aligned}R_{2n+1}^{(3)}(\nu,\mu)\tilde{\mathbf{Q}}_\nu^{i\mu}(x)&\\ =\left(\frac{\alpha^2-\zeta}{1+\alpha^2-x^2}\right)^{1/4}&\left[G_{iu\alpha}(u|\zeta|^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{u^{2s}}+\frac{|\zeta|^{1/2}}{u}G'_{iu\alpha}(u|\zeta|^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{u^{2s}}\right.\\ &\left.+\frac{i}{\pi|\zeta|^{1/2}}\{\varepsilon_{2n+1}^{(3)}(u,\alpha,\zeta+i0)-\varepsilon_{2n+1}^{(3)}(u,\alpha,\zeta-i0)\}\right].\end{aligned} \tag{6.8}$$

We remark that, since the LHS of this equation is real, the error term on the RHS must also be real. However, unlike the corresponding result in [1, eq. (6.12)], it does not follow from their integral equations that $\varepsilon_{2n+1}^{(3)}(u, \alpha, \zeta + i0)$ and $\varepsilon_{2n+1}^{(3)}(u, \alpha, \zeta - i0)$ are necessarily complex conjugates.

REFERENCES

[1] W. G. C. BOYD AND T. M. DUNSTER, *Uniform asymptotic solutions of a class of second-order linear differential equations having a turning point and regular singularity, with an application to Legendre functions*, SIAM J. Math. Anal., 17 (1986), pp. 422–450.

[2] T. M. DUNSTER, *Uniform asymptotic expansions for prolate spheroidal functions with large parameters*, SIAM J. Math. Anal., 17 (1986), pp. 1495–1524.

[3] ———, *Uniform asymptotic expansions for Whittaker's confluent hypergeometric functions*, SIAM J. Math. Anal., 20 (1989), pp. 744–760.

[4] ———, *Bessel functions of purely imaginary order, with an application to second-order linear differential equations having a large parameter*, SIAM J. Math. Anal., 21 (1990), pp. 994–1017.

[5] J. J. NESTOR, *Uniform asymptotic approximations of solutions of second-order linear differential equations, with a coalescing simple turning point and simple pole*, Ph.D. thesis, University of Maryland, College Park, MD, 1984.

[6] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[7] ———, *Second-order linear differential equations with two turning points*, Philos. Trans. Roy. Soc. London Ser. A, 278 (1975), pp. 137–174.

[8] ———, *Legendre functions with both parameters large*, Philos. Trans. Roy. Soc. London Ser. A, 278 (1975), pp. 175–185.

[9] ———, *Whittaker functions with both parameters large: uniform approximations in terms of parabolic cylinder functions*, Proc. Roy. Soc. Edinburgh Sect. A (1980), pp. 213–234.

[10] ———, *Asymptotic expansions and error bounds*, SIAM Rev., 22 (1980), pp. 188–203.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1619-1632, November 1990
© 1990 Society for Industrial and Applied Mathematics
013

A UNIFORM EXPANSION FOR THE EIGENFUNCTION OF A SINGULAR SECOND-ORDER DIFFERENTIAL OPERATOR*

A. FITOUHI† AND M. M. HAMZA†

Abstract. In a recent work, Frenzen and Wong [*Canad. J. Math.*, 37 (1985), pp. 979-1007] have obtained a uniform asymptotic expansion for the Jacobi polynomials in terms of Bessel functions. An analogous expansion for the Jacobi functions had been given earlier by Stanton and Tomas [*Acta Math.*, 140 (1978), pp. 251-276]. The common starting point of these papers is an integral representation.

In this paper it is shown that in general such expansions can be obtained directly from an eigenfunction of a singular second-order differential operator, and with additional assumptions they converge in some interval. This leads to an expansion for the eigenfunction of an integral representation of Mehler type with good information on the kernel.

Key words. singular second-order differential equation, Bessel functions, Jacobi functions, integral equation, asymptotic expansion

AMS(MOS) subject classifications. 33A65, 33A40, 34E05

1. Introduction. In [3] Frenzen and Wong have established an asymptotic expansion for Jacobi polynomials $P_n^{(\alpha,\beta)}(\cos\theta)$ in terms of Bessel functions $J_{\alpha+p}$, $p=0, 1, 2, \cdots$. They used an integral representation of the Jacobi polynomials due to Gasper [4].

Stanton and Tomas [8] have given a similar expansion for spherical functions on noncompact symmetric spaces; they used an integral representation due to Koornwinder [5].

We note that such expansions can be rewritten as expansions of Olver type [7, Chap. 12] for more general differential operators.

In this paper we are concerned with a class of second-order singular differential equations of the form

$$\text{(A)} \qquad u''-\left(\frac{\alpha^2-\frac{1}{4}}{x^2}-\mu^2-\chi(x)\right)u(x)=0, \qquad \alpha>-\frac{1}{2}.$$

This class contains (after simple transformation) a wide class of differential equations of the form

$$\frac{1}{x^{2\alpha+1}C(x)}[x^{2\alpha+1}C(x)u'(x)]'+(\mu^2+q(x))u(x)=0$$

with $C(x)$ an even positive function which is analytic or $\mathscr{C}^\infty$. The radial parts of the Laplace-Beltrami operator on a symmetric Riemannian space of rank one are of this type; the differential equation related to special functions such as those of Legendre, Gegenbauer, and Jacobi on a finite or an unbounded interval are also of this type.

Following Olver [7] we begin by seeking a formal series solution of equation (A) of the form

$$V_\mu(x)=\sum_{p=0}^{\infty} x^{p+1/2}B_p(x)\frac{J_{\alpha+p}(\mu x)}{\mu^{\alpha+p}}.$$

* Received by the editors November 10, 1986; accepted for publication (in revised form) December 11, 1989.

† Département de Mathématiques, Faculté des Sciences de Tunis, Campus Universitaire, 1060 Tunis, Tunisie.

The central result is that this series is a solution of (A) on some interval of the real line when $\chi(z)$ is an even holomorphic function in a disc centered at the origin. This result leads to an integral representation of Mehler type for the solution with a kernel

$$k(x,t)=(1-t^2)^{\alpha-1/2}\sum_{p=0}^{\infty}\left(\frac{x^2}{2}\right)^p\frac{B_p(x)}{\Gamma(\alpha+p+\frac{1}{2})}(1-t^2)^p.$$

We note that Trimèche [9, p. 71] has combined an asymptotic expansion of Olver type and a Paley-Wiener theorem to obtain a similar integral representation but with less information on the kernel.

In the case $\chi(x)\subset\mathscr{C}^2$, by using the results and the methods of Olver [7, Chap. 12] we deduce the following estimate:

$$|\mathscr{R}_{m,\mu}(x)|\leqq\frac{\mathscr{E}_m(x)}{|\mu|^{\alpha+m+3/2}}$$

with an upper bound of the majorant $\mathscr{E}_m(x)$, where $\mathscr{R}_{m,\mu}(x)$ is the remainder term defined by

$$\mathscr{R}_{m,\mu}(x)=V_\mu(x)-\sum_{p=0}^{m}x^{p+1/2}B_p(x)\frac{J_{\alpha+p}(\mu x)}{\mu^{\alpha+p}},\quad \mu \text{ real}.$$

The methods are more constructive and direct than those of [3] and [8] and lead to more general results. For instance, we obtain the asymptotic expansion for the Jacobi polynomials $P_n^{(\alpha,\beta)}(\cos\theta)$ given in [3] with only the restrictions $\alpha>-\frac{1}{2}$, $\beta>-1$. We also obtain the asymptotic expansion for the Jacobi functions when α and β are half odd integers given in [8], with only the restriction $\alpha>-\frac{1}{2}$.

2. Formal computation. Let V_μ be a solution of the differential equation

$$\frac{d^2V_\mu}{dx^2}=\left\{\frac{\alpha^2-\frac{1}{4}}{x^2}-\mu^2-\chi(x)\right\}V_\mu,\qquad \alpha>-\frac{1}{2},\tag{2.1}$$

when $x\in(0,a)$. Here a is a positive constant, μ is a real or a complex parameter, and $\chi(x)$ is an even sufficiently differentiable function.

We shall seek a formal solution of (2.1) of the form

$$\sum_{p=0}^{\infty}A_p(x)\sqrt{x}\frac{J_{\alpha+p}(\mu x)}{\mu^{\alpha+p}},$$

where $J_\nu(x)$ denotes the Bessel function of first kind of index ν. We recall that $\sqrt{x}\,J_\nu(\mu x)$ is a solution of the equation

$$u''=\left(\frac{\nu^2-\frac{1}{4}}{x^2}-\mu^2\right)u,$$

and that the functions $J_\nu(x)$ satisfy the recurrence relations

$$J_{\nu-1}(x)+J_{\nu+1}(x)=\frac{2\nu}{x}J_\nu(x),$$

$$J_{\nu-1}(x)-J_{\nu+1}(x)=2J'_\nu(x).$$

We denote by L_α the operator defined by

$$L_\alpha u=u''-\frac{\alpha^2-\frac{1}{4}}{x^2}u.$$

Using these relations, we find that

$$(L_\alpha+\mu^2+\chi(x))(A_p(x)\sqrt{x}\,J_{\alpha+p}(\mu x))$$
$$=\sqrt{x}\,J_{\alpha+p}(\mu x)\left\{A_p''(x)+\frac{1-2(\alpha+p)}{x}A_p'(x)+\frac{p(p+2\alpha)}{x^2}A_p(x)+\chi(x)A_p(x)\right\}$$
$$+2\mu\sqrt{x}\,J_{\alpha+p-1}(\mu x)A_p'(x).$$

Thus if the series $\sum_0^\infty A_p(x)(\sqrt{x}\,J_{\alpha+p}(\mu x)/\mu^{\alpha+p})$ is to be a formal solution of (2.1) it suffices to choose the coefficients $A_p(x)$ to satisfy the following relations:

$$A_0'(x)=0,$$
$$-2A_{p+1}'(x)=A_p''(x)+\frac{1-2(\alpha+p)}{x}A_p'(x)+\left\{\frac{p(p+2\alpha)}{x^2}+\chi(x)\right\}A_p(x). \tag{2.2}$$

In what follows we show that such a choice is possible and can be made to yield coefficients $A_p(x)$ that are analytic or $\mathscr{C}^\infty$ functions according to whether the function $\chi(x)$ is analytic or $\mathscr{C}^\infty$.

Let us put $A_p(x)=x^pB_p(x)$. Then (2.2) can be written as

$$B_0'(x)=0,$$
$$\{x^{p+1}B_{p+1}(x)\}'=-\frac{1}{2}x^p\left\{B_p''(x)+\frac{1-2\alpha}{x}B_p'(x)+\chi(x)B_p(x)\right\}. \tag{2.3}$$

Denoting by B_0 the constant value of $B_0(x)$, we choose $B_1(x)$ by

$$B_1(x)=\begin{cases}-\dfrac{B_0}{2x}\displaystyle\int_0^x\chi(t)\,dt & \text{if } x\neq 0,\\[2ex] -\dfrac{B_0}{2}\chi(0) & \text{if } x=0.\end{cases}$$

It is clear that $B_1(x)$ is an even analytic or $\mathscr{C}^\infty$ function depending on the nature of $\chi(x)$. From the relations

$$B_{p+1}(x)=\begin{cases}-\dfrac{1}{2x^{p+1}}\displaystyle\int_0^x t^p\left\{B_p''(t)+\frac{1-2\alpha}{t}B_p'(t)+\chi(t)B_p(t)\right\}dt & \text{if } x\neq 0,\\[2ex] -\dfrac{1}{2(p+1)}\{2(1-\alpha)B_p''(0)+\chi(0)B_p(0)\} & \text{if } x=0,\end{cases} \tag{2.4}$$

we see by induction that the same is true of each functoin $B_p(x)$. Thus we have the formal solution

$$V_\mu(x)=\sum_{p=0}^\infty x^pB_p(x)\frac{\sqrt{x}\,J_{\alpha+p}(\mu x)}{\mu^{\alpha+p}} \tag{2.5}$$

of (2.1) where the functions $B_p(x)$ are defined by the recursive formula (2.4) and are even and analytic or $\mathscr{C}^\infty$, depending on the nature of $\chi(x)$.

3. Convergence of the series (2.5). In this section we show that with the additional assumption that $\chi(z)$ is holomorphic in a disc $D(0,2b)=\{z\in\mathbb{C},|z|<2b\}$, the series (2.5) converges on an interval of the real line. This is achieved by estimating the coefficients $B_p(x)$ by using Leibniz's formula and the maximum-modulus theorem.

THEOREM 1. *Suppose that $\chi(z)$ is holomorphic in the disc $D(0,2b)$. Then the functions $B_p(x)$ defined by the recursive relation* (2.3) *satisfy*

$$|B_p^{(q)}(x)| \leqq (c/2)^p d^{p-1} b^{1-p-q} (p+q-1)! \tag{3.1}$$

for all $x \in [0, b]$, $p = 1, 2, \cdots$ and $q = 0, 1, 2, \cdots$, where

$$c = \max\{1 + |2\alpha - 1|, 2 \sup(|\chi(z)|, |z| \leqq 2b)\}, \qquad d = (b + b^{-1}).$$

Proof. Since $\chi(z)$ is holomorphic in the disc $D(0, 2b)$, by the maximum-modulus theorem we have

$$|\chi^{(q)}(x)| \leqq \beta b^{-q} q!, \qquad q = 0, 1, 2, \cdots, \quad x \in [0, b], \tag{3.2}$$

where $\beta = 2 \sup\{|\chi(z)|, |z| \leqq 2b\}$. If f is a $\mathscr{C}^\infty$ function on $\mathbb{R}$, then for $p = 1, 2, \cdots$ let us denote by $H_p f$ the function defined on $\mathbb{R}$ by

$$(H_p f)(x) = \begin{cases} \dfrac{1}{x^p} \displaystyle\int_0^x t^{p-1} f(t)\, dt & \text{if } x \neq 0, \\ \dfrac{1}{p} f(0) & \text{if } x = 0. \end{cases} \tag{3.3}$$

By Lemma 3.1 of [9], $H_p f$ is a $\mathscr{C}^\infty$ function on $\mathbb{R}$ and

$$(H_p f)^{(q)} = H_{p+q}(f^{(q)}), \qquad q = 0, 1, 2, \cdots.$$

Thus the recursive formula (2.4) can be written as

$$B_{p+1} = -\tfrac{1}{2} H_{p+1}[B_p'' + (1 - 2\alpha) H_1 B_p'' + \chi B_p],$$

and hence we obtain

$$B_{p+1}^{(q)} = -\tfrac{1}{2} H_{p+q+1}[B_p^{(q+2)} + (1 - 2\alpha) H_{q+1} B_p^{(q+2)} + (\chi B_p)^{(q)}].$$

This relation leads to the following estimate:

$$|B_{p+1}^{(q)}|_\infty \leqq \frac{1}{2(p+q+1)} \{(1 + |2\alpha - 1|)|B_p^{(q+2)}|_\infty + |(\chi B_p)^{(q)}|_\infty\}, \tag{3.4}$$

where

$$|u|_\infty = \sup\{|u(x)|,\ 0 < x < b\}.$$

Using Leibniz's formula and the estimate (3.2) we deduce that

$$|(\chi B_p)^{(q)}|_\infty \leqq \beta \frac{q!}{b^q} \sum_{j=0}^{q} \frac{b^j}{j!} |B_p^{(j)}|_\infty. \tag{3.5}$$

Substituting (3.5) in (3.4) we obtain

$$|B_{p+1}^{(q)}| \leqq \frac{c}{2(p+q+1)} \left\{ |B_p^{(q+2)}|_\infty + \frac{q!}{b^q} \sum_{j=0}^{q} \frac{b^j}{j!} |B_p^{(j)}|_\infty \right\}, \tag{3.6}$$

where $c = \max(1 + |2\alpha - 1|, \beta)$. Using (3.6) we shall prove (3.1) by induction. Since $B_1 = -\frac{1}{2} H_1 \chi$, we have $B_1^{(q)} = -\frac{1}{2} H_{q+1} \chi^{(q)}$. From (3.2) we deduce that

$$|B_1^{(q)}| \leqq \frac{c}{2(q+1)} q! \left(\frac{1}{b}\right)^q \leqq \frac{c}{2} q! \left(\frac{1}{b}\right)^q.$$

If $p=1$, then the recurrence hypothesis is true for all integers q. Suppose that (3.1) is valid for a given positive integer p, and for all integers q. Then using (3.6) we have

$$|B_{p+1}^{(q)}|_\infty \leqq \left(\frac{c}{2}\right)^{p+1}\left(\frac{1}{b}\right)^{p+q}(q+p)!\,d^{p-1}\left\{\frac{1}{b}+b\frac{q!}{(q+1+p)!}\sum_{j=0}^{q}\frac{(j+p-1)!}{j!}\right\},$$

and by noting that the sequence $\{(j+p-1)!/j!\}$ is increasing we deduce that

$$\frac{q!}{(q+p+1)!}\sum_{j=0}^{q}\frac{(j+p-1)!}{j!}\leqq\frac{(q+1)!}{(q+p+1)!}\frac{(q+p-1)!}{q!}\leqq 1.$$

The required result is a consequence of this inequality.

THEOREM 2. *With the assumptions of Theorem* 1 *and the conditions*

$$V_\mu(x)\sim x^{\alpha+1/2},\quad V'_\mu(x)\sim(\alpha+\tfrac{1}{2})x^{\alpha-1/2}\qquad(x\to 0),\tag{3.7}$$

equation (2.1) *has a unique solution* $V_\mu(x)$ *such that*

$$V_\mu(x)=\sum_{p=0}^{\infty}x^pB_p(x)\frac{\sqrt{x}\,J_{\alpha+p}(\mu x)}{\mu^{\alpha+p}},\tag{3.8}$$

the infinite series being uniformly convergent on $[0,\gamma]$ *with* $\gamma=\min\{b,2(b/(cd))^{1/2}\}$.

Proof. From (3.1) and the fact that for real x and complex μ

$$|J_{\alpha+p}(\mu x)|\leqq e^{|\mu x|}\frac{|\mu x|^{\alpha+p}}{2^{\alpha+p}\Gamma(\alpha+p+1)},$$

we obtain

$$\left|x^pB_p(x)\frac{\sqrt{x}\,J_{\alpha+p}(\mu x)}{\mu^{\alpha+p}}\right|\leqq e^{|\mu x|}\frac{|x|^{\alpha+1/2}}{2^\alpha}\frac{b}{d}\left(c\frac{dx^2}{4b}\right)^p\frac{\Gamma(p)}{\Gamma(p+\alpha+1)}.$$

The right member of this inequality is the general term of a uniformly convergent series on $[0,\gamma]$ with γ defined as above.

THEOREM 3. *For* $x\in[0,\gamma]$, *where* γ *is as defined in Theorem* 2, *the solution* V_μ *of* (2.1) *that satisfies the conditions* (3.7) *has the following representation of Mehler type*:

$$V_\mu(x)=\frac{1}{\sqrt{\pi}}\int_{-1}^{1}K(x,t)\cos(\mu xt)\,dt,$$

with

$$K(x,t)=(1-t^2)^{\alpha-1/2}H(x,t),$$

and

$$H(x,t)=\sum_{p=0}^{\infty}\left(\frac{x^2}{2}\right)^p\frac{B_p(x)}{\Gamma(\alpha+p+\frac{1}{2})}(1-t^2)^p.$$

Proof. It is known that [6, p. 114]

$$J_\nu(z)=\frac{(z/2)^\nu}{\sqrt{\pi}\,\Gamma(\nu+\frac{1}{2})}\int_{-1}^{1}(1-t^2)^{\nu-1/2}\cos(zt)\,dt.$$

From Theorem 2 for $x\in[0,\gamma)$ we have

$$V_\mu(x)=\frac{x^{\alpha+1/2}}{\sqrt{\pi}}\sum_{p=0}^{\infty}\frac{(x^2/2)^pB_p(x)}{\Gamma(\alpha+p+\frac{1}{2})}\int_{-1}^{1}(1-t^2)^{\alpha+p-1/2}\cos(\mu xt)\,dt.$$

On the other hand, for $t \in [0, 1]$ we have

$$\left|\left(\frac{x^2}{2}\right)^p \frac{B_p(x)}{\Gamma(\alpha+p+\frac{1}{2})}(1-t^2)^{\alpha+p-1/2}\cos(\mu x t)\right| \leqq e^{|\mu x|}\left|\frac{x^{2p}}{2^p}\frac{B_p(x)}{\Gamma(\alpha+p+\frac{1}{2})}\right|.$$

The second member of this inequality is the general term of a convergent series on $[0, \gamma)$, thus the interchange of the operations $\int, \sum$ is valid, and the result follows.

4. Asymptotic expansion of V_μ for real μ. Let us write

$$V_\mu(x) = \sum_{p=0}^{m} A_p(x)\frac{\sqrt{x}\,J_{\alpha+p}(\mu x)}{\mu^{\alpha+p}} + \mathcal{R}_{m,\mu}(x).$$

In [7, p. 445] a majorization of $|\mathcal{R}_{m,\mu}(x)|$ is given for a more general equation than (2.1). By taking this into account we are able to give an estimate for $\mathcal{R}_{m,\mu}(x)$, which improves and generalizes the results of [3] and [8].

Using $(L_\alpha + \mu^2 + \chi)V_\mu = 0$ and the relation (2.2), we obtain

$$(L_\alpha + \mu^2)\mathcal{R}_{m,\mu}(x) = -\chi(x)\mathcal{R}_{m,\mu}(x) + 2A'_{m+1}(x)\sqrt{x}\frac{J_{\alpha+m}(\mu x)}{\mu^{\alpha+m}}.$$

The functions $\sqrt{x}\,J_\alpha(\mu x)$ and $\sqrt{x}\,Y_\alpha(\mu x)$, where $Y_\nu(x)$ denotes the Bessel function of second kind, are two linearly independent solutions of the equation $(L_\alpha + \mu^2)u = 0$; thus the method of variation of parameters shows that $\mathcal{R}_{m,\mu}$ is a solution of the following singular Volterra integral equation:

$$y(x) = \int_0^x \frac{\pi}{2}\sqrt{xt}\left[\chi(t)y(t) - 2A'_{m+1}(t)\frac{\sqrt{t}\,J_{\alpha+m}(\mu t)}{\mu^{\alpha+m}}\right]$$
$$\cdot\,[Y_\alpha(\mu t)J_\alpha(\mu x) - J_\alpha(\mu t)Y_\alpha(\mu x)]\,dt.$$

This equation has the form

$$y(x) = \int_0^x H_\alpha(\mu x, \mu t)[\varphi(t)g(\mu t) + \psi(t)y(t)]\,dt,$$

where the functions H_α, ψ, φ, and g are defined by

$$H_\alpha(x, t) = \left(\frac{\pi}{2}\right)\sqrt{xt}[J_\alpha(x)Y_\alpha(t) - J_\alpha(t)Y_\alpha(x)],$$

$$\varphi(t)g(\mu t) = \frac{-2t}{\mu^{\alpha+m+1}}J_{\alpha+m}(\mu t)A'_{m+1}(t),$$

$$\psi(t) = \frac{1}{\mu}\chi(t).$$

To deduce an integral inequality satisfied by the remainder $\mathcal{R}_{m,\mu}$ we recall the following estimate of the kernel $H_\alpha(x, t)$ from [7, pp. 437, 445].

Hypothesis. Suppose that

$$|H_\alpha(x, t)| \leqq \frac{\pi}{2}P_\alpha(x)Q_\alpha(t),$$

where $P_\alpha(x)$ and $Q_\alpha(t)$ are the (continuous) functions defined by

(i) If $-\frac{1}{2}<\alpha\leqq\frac{1}{2}$, $P_\alpha(x)=Q_\alpha(x)=\sqrt{2/\pi}$,

(ii) If $\alpha>\frac{1}{2}$

$$P_\alpha(x)=\begin{cases}\sqrt{2x}\,|J_\alpha(x)| & \text{if } 0<x\leqq X_\alpha,\\ \sqrt{N_\alpha(x)} & \text{if } x\geqq X_\alpha,\end{cases}$$

$$Q_\alpha(x)=\begin{cases}\sqrt{2x}\,|Y_\alpha(x)| & \text{if } 0<x\leqq X_\alpha,\\ \sqrt{N_\alpha(x)} & \text{if } x\geqq X_\alpha,\end{cases}$$

where $N_\alpha(x)=x(J_\alpha^2(x)+Y_\alpha^2(x))$ is the Nicolson function, and X_α is the first positive zero of $J_\alpha(x)+Y_\alpha(x)$.

By application of these definitions, we see that the remainder $\mathcal{R}_{m,\mu}$ satisfies the following integral inequality

$$|y(x)|\leqq\frac{\pi}{2}P_\alpha(\mu x)\int_0^x Q_\alpha(\mu t)\left\{|\varphi(t)g(\mu t)|+\left|\frac{\chi(t)}{\mu}\right||y(t)|\right\}dt, \tag{4.1}$$

where

$$\varphi(t)=-2\mu^{-(\alpha+m+3/2)}A'_{m+1}(t),\qquad g(\mu t)=\sqrt{\mu t}\,J_{\alpha+m}(\mu t), \tag{4.2}$$

or

$$\varphi(t)=-2\mu^{-(\alpha+m+1)}\sqrt{t}\,A'_{m+1}(t),\qquad g(\mu t)=J_{\alpha+m}(\mu t). \tag{4.3}$$

Since $P_\alpha(x)$, $P_\alpha(x)Q_\alpha(x)$, and $\sqrt{x}\,J_{\alpha+m}(x)Q_\alpha(x)$ are continuous on $[0,\infty)$ and have finite limits at zero and infinity, the following suprema are finite:

$$\begin{aligned}k_0(\alpha)&=\sup\{P_\alpha(x)Q_\alpha(x),\,x>0\},\\ k_1(\alpha)&=\sup\{P_\alpha(x),\,x>0\},\\ c_m(\alpha)&=\sup\{\sqrt{x}\,|J_{\alpha+m}(x)Q_\alpha(x)|,\,x>0\}.\end{aligned} \tag{4.4}$$

Note that $\pi k_0(\alpha)$ is denoted in [7, p. 443] by $\lambda_2(\alpha)$ and $\pi c_0(\alpha)$ by $\lambda_3(\alpha)$. By applying Theorem 10.1 of [7, p. 219] with φ and g given by (4.2) we obtain the next theorem.

THEOREM 4. *Let $\chi(x)\subset\mathcal{C}^2$ and V_μ be the unique solution of* (2.1) *that satisfies the conditions* (3.7). *Then*

$$V_\mu(x)=\sum_{p=0}^{m}x^{p+1/2}B_p(x)\frac{J_{\alpha+p}(\mu x)}{\mu^{\alpha+p}}+\mathcal{R}_{m,\mu}(x), \tag{4.5}$$

where the coefficients $B_p(x)$ are defined by the recursive relation (2.4) *and the remainder $\mathcal{R}_{m,\mu}$ satisfies*

$$|\mathcal{R}_{m,\mu}(x)|\leqq\frac{1}{|\mu|^{\alpha+m+3/2}}\mathcal{E}_{m,\mu}(x)\exp\left(k_0(\alpha)\int_0^x\frac{|\chi(t)|}{|\mu|}dt\right), \tag{4.6}$$

with

$$\mathcal{E}_{m,\mu}(x)=\pi P_\alpha(\mu x)\left\{\sup_{0\leqq t\leqq\mu x}|\sqrt{t}\,J_{\alpha+m}(t)Q_\alpha(t)|\right\}\int_0^x|(t^{m+1}B_{m+1}(t))'|\,dt.$$

Furthermore,

$$\mathcal{E}_{m,\mu}(x)=\pi k_1(\alpha)c_m(\alpha)\int_0^x|(t^{m+1}B_{m+1}(t))'|\,dt,$$

where $k_1(\alpha)$ and $c_m(\alpha)$ are given by (4.4).

Remark. (1) In Theorem 2 we established that the series (2.5) is convergent on some interval of the real line. When we are outside this interval, or when the function χ is $\mathscr{C}^2$ and does not satisfy the assumptions of Theorem 2, Theorem 4 shows that the series (2.5) is a useful approximation of V_μ for large μ.

(2) Under the assumptions of Theorem 2, and with φ and g given by (4.3), it can be shown that

$$|\mathscr{R}_{m,\mu}(x)| \leqq k \frac{d^{m+1}}{(2b)^m} \frac{m!}{\mu^{\alpha+m+1}} x^{m+3/2} \quad \text{on } [0, b], \quad \mu \geqq 1,$$

where the constant k depends only on χ and b.

(3) In addition, since $\mathscr{R}_{m,\mu}$ satisfies the integral inequality (4.1) it can be shown by using Gronwall's lemma, rather than Picard's method of successive approximations, that $\mathscr{E}_{m,\mu}(x)$ can be taken to be

$$\mathscr{E}_{m,\mu}(x) = \pi P_\alpha(\mu x) \int_0^x \sqrt{\mu t}\, |J_{\alpha+m}(\mu t) Q_\alpha(\mu t) \{t^{m+1} B_{m+1}(t)\}'|\, dt.$$

5. Error bounds. In this section we give upper bounds for the suprema defined above.

LEMMA. *The Bessel functions satisfy*

(i) $|J_\nu(x)| \leqq \dfrac{1}{2^\nu \Gamma(\nu+1)} x^\nu, \qquad x>0, \quad \nu > -\dfrac{1}{2};$

(ii) $|J_{\nu+m}(x)| \leqq \dfrac{\sqrt{\pi}}{2^\nu \Gamma(\nu+\frac{1}{2})} x^\nu, \quad x>0, \quad \nu>0, \quad m \in \mathbb{N};$

(iii) $|J_\nu(x)| \leqq \dfrac{1}{2} + \dfrac{1}{\nu\pi}, \qquad x>0, \quad \nu>0;$

(iv) $|x^\nu H_\nu^{(1)}(x)| \leqq \sqrt{\dfrac{2}{\pi}} \left\{ (2x)^{\nu-1/2} + \dfrac{2^{2\nu-1}}{\sqrt{\pi}} \Gamma(\nu) \right\}, \qquad x>0, \quad \nu > \dfrac{1}{2};$

(v) $|H_\nu^{(1)}(x)| \leqq \sqrt{\dfrac{2}{\pi x}} \left\{ 2^{\nu-1/2} + \dfrac{2^{2\nu-1}}{\sqrt{\pi}\, x^{\nu+1/2}} \Gamma(\nu+1) \right\}, \qquad x>0, \quad \nu > \dfrac{1}{2}.$

Here $H_\nu^{(1)}(x) = J_\nu(x) + iY_\nu(x)$ is the Hankel function in the usual notation.

Inequalities (i) and (iii) are consequences of the integral representation [11, p. 48] and [11, p. 177]. The use of the Sonine formula [11, p. 373] leads to the inequality (ii). To prove (iv) and (v) we use the integral representation of $H_\nu^{(1)}(x)$ given in [11, p. 168].

For (iv) we subdivide the interval $[0, \infty)$ into $[0, 2x]$ and $[2x, \infty)$, and bound $1 + t/2x$ by 2 in the first part and by t/x in the second part.

For (v) we use

$$\left|1 + \frac{t}{2x}\right|^{\nu-1/2} \leqq \left|1 + \frac{t}{2x}\right|^{\nu+1/2}$$

and the convexity of the function $u^{\nu+1/2}$.

5.1. Case $-\frac{1}{2} < \alpha < \frac{1}{2}$. For $m = 1, 2, \cdots$ we use the inequalities (iii) and (v) of the lemma to bound $\sqrt{x}\, |J_{\alpha+m}(x)|$, respectively, on $[0, 1]$ and $[1, \infty)$ to obtain

$$c_m(\alpha) \leqq \max \left\{ \frac{1}{2} + \frac{1}{(\alpha+1)\pi}, \frac{2^{\alpha+m}}{\sqrt{\pi}} + \frac{2^{2\alpha+2m}}{\sqrt{2\pi}} \Gamma(\alpha+m+1) \right\}.$$

Furthermore, it is clear that $c_0(\alpha) \leqq \sqrt{2/\pi}$.

5.2. Case $\alpha > \frac{1}{2}$. In this case, the Nicolson function $N_\alpha(x)$ is continuous and decreasing [11, p. 416]. In consequence, the suprema $k_0(\alpha)$ and $k_1(\alpha)$ are attained on $(0, X_\alpha]$, and we deduce by (i) and (v) of the lemma that

$$k_0(\alpha) \leqq \frac{\sqrt{\pi}}{\Gamma(\alpha+1)} \left\{ X_\alpha^{\alpha+1/2} + \frac{2^{\alpha-1/2}}{\sqrt{\pi}} \Gamma(\alpha+1) \right\},$$

and by (iii) of the lemma that

$$2\alpha k_1(\alpha) \leqq (2+\alpha\pi) \left(\frac{X_\alpha}{2} \right)^{1/2}.$$

Using first (ii) and then (iv) of the lemma we obtain, for $0 < x \leqq X_\alpha$,

$$\begin{aligned} \sqrt{x}|J_{\alpha+m}(x)Q_\alpha(x)| &\leqq \frac{\sqrt{2\pi}}{2^\alpha \Gamma(\alpha+\frac{1}{2})} x^{\alpha+1} |Y_\alpha(x)| \\ &\leqq \frac{X_\alpha}{\Gamma(\alpha+\frac{1}{2})} \left\{ \frac{(X_\alpha)^{\alpha-1/2}}{\sqrt{2}} + \frac{2^\alpha}{\sqrt{\pi}} \Gamma(\alpha) \right\}. \end{aligned}$$

We have also by (v) of the lemma and the inequality

$$|Q_\alpha(x)| \leqq \sqrt{2X_\alpha}\, J_\alpha(X_\alpha), \qquad x \geqq X_\alpha,$$

$$\sqrt{x}|J_{\alpha+m}(x)Q_\alpha(x)| \leqq \sqrt{X_\alpha}\, J_\alpha(X_\alpha) \left\{ \frac{2^{\alpha+m+1/2}}{\sqrt{\pi}} + \frac{2^{2\alpha+2m}}{\pi (X_\alpha)^{\alpha+m+1/2}} + \Gamma(\alpha+m+1) \right\}, \quad x \geqq X_\alpha.$$

Finally, we have

$$\begin{aligned} c_m(\alpha) \leqq \max \Bigg[& X_\alpha \frac{\sqrt{\pi}(X_\alpha)^{\alpha-1/2} + 2^{\alpha+1/2}\Gamma(\alpha)}{\sqrt{2\pi}\,\Gamma(\alpha+\frac{1}{2})}, \\ & \pi^{-1}(X_\alpha)^{-\alpha-m} J_\alpha(X_\alpha) \left(\sqrt{\pi} \left(\frac{2}{X_\alpha} \right)^{\alpha+m+1/2} + 2^{2\alpha+2m}\Gamma(\alpha+m+1) \right) \Bigg]. \end{aligned}$$

6. Applications.

6.1. Case $\chi(x)$ is constant. In (2.1) if $\chi(x) = \lambda$, where λ is a given real or complex number, then the coefficients $B_p(x)$ defined by (2.4) are given by

$$B_p(x) = \left(\frac{-\lambda}{2} \right)^p \left(\frac{1}{p!} \right), \qquad p \in \mathbb{N},$$

and it is clear then that the series

$$\sum_{p=0}^{\infty} \frac{1}{p!} \left(\frac{-\lambda x}{2} \right)^p \frac{\sqrt{x}\, J_{\alpha+p}(\mu x)}{\mu^{\alpha+p}}$$

is uniformly convergent on every compact interval of $[0, \infty)$, and is the unique solution of (2.1) that satisfies the conditions (3.7).

From these considerations we deduce the following classical relations:

(i) With $\chi(x) = \rho^2$, $\rho \in \mathbb{C}$, we obtain

$$(\mu^2+\rho^2)^{-\alpha/2} J_\alpha(\sqrt{\mu^2+\rho^2}\, x) = \sum_{p=0}^{\infty} \frac{1}{p!} \left(-\frac{\rho^2 x}{2} \right)^p \frac{J_{\alpha+p}(\mu x)}{\mu^{\alpha+p}}, \qquad x > 0.$$

(ii) With $\chi(x) = \rho^2 - \mu^2$, $\rho \in \mathbb{C}$, and $\alpha = \frac{1}{2}$, we obtain

$$\sqrt{\frac{1}{\pi}} \frac{\sin \rho x}{\rho} = \sum_{p=0}^{\infty} \frac{1}{p!} (\mu^2-\rho^2)^p \left(\frac{x}{2\mu} \right)^{p+1/2} J_{p+1/2}(\mu x), \qquad x > 0.$$

6.2. Uniform asymptotic expansion for the eigenfunction of a singular second-order differential operator. With $a \in (0, \infty)$ we denote by L a differential operator of the second order on $(0, a)$ having the form

$$(Lu)(x) = \frac{1}{x^{2\alpha+1}C(x)} \{x^{2\alpha+1}C(x)u'(x)\}' + q(x)u(x),$$

where $\alpha > -\frac{1}{2}$, and the functions $C(x)$ and $q(x)$ are even and analytic or $\mathscr{C}^\infty$; furthermore, $C(x) > 0$.

These operators generalize second-order differential operators related to special functions. The Bessel operator is obtained by putting $a = +\infty$, $C(x) = 1$, and $q(x) = 0$. The Jacobi operators on $(0, \pi)$ or $(0, \infty)$ are obtained by taking, respectively,

$$\begin{aligned} C(x) &= \left(\frac{\sin(x/2)}{x}\right)^{2\alpha+1} \left(\cos\frac{x}{2}\right)^{2\beta+1}, \qquad 0 < x < \pi, \\ q(x) &= -\frac{1}{4}(\alpha+\beta+1)^2 \end{aligned} \tag{6.1}$$

and

$$\begin{aligned} C(x) &= 2^{2(\alpha+\beta+1)} \left(\frac{\operatorname{sh} x}{x}\right)^{2\alpha+1} (\operatorname{ch} x)^{2\beta+1}, \qquad 0 < x < \infty, \\ q(x) &= (\alpha+\beta+1)^2. \end{aligned} \tag{6.2}$$

When $\alpha = \beta$ we have the Gegenbauer operator. The Legendre operator is obtained by setting $\alpha = \beta = 0$. In what follows we consider a real or complex parameter μ for which it is known [2] that there exists a unique function φ_μ satisfying

$$\begin{aligned} &L\varphi_\mu + \mu^2\varphi_\mu = 0 \quad \text{on } (0, a), \\ &\varphi_\mu(0) = 1, \qquad \varphi'_\mu(0) = 0. \end{aligned} \tag{6.3}$$

Putting $V_\mu(x) = x^{\alpha+1/2}\sqrt{C(x)}\,\varphi_\mu(x)$, we see that V_μ satisfies

$$(L_\alpha + \chi(x) + \mu^2)V_\mu(x) = 0,$$

where L_α, the Bessel operator, and $\chi(x)$ are defined by

$$\begin{aligned} (L_\alpha u)(x) &= u''(x) - \frac{\alpha^2 - \frac{1}{4}}{x^2} u(x), \\ \chi(x) &= q(x) - (2\alpha+1)\frac{C'(x)}{2xC(x)} - \frac{1}{2}\left(\frac{C'(x)}{C(x)}\right)' - \frac{1}{4}\left(\frac{C'(x)}{C(x)}\right)^2. \end{aligned} \tag{6.4}$$

In [9, p. 60] Trimèche gives the following asymptotic expansion of Olver type [7, Chap. 12] for the solution φ_μ of (6.3), μ being real and $n = 1, 2, \cdots$:

$$\sqrt{\frac{C(x)}{C(0)}}\,\varphi_\mu(x) = j_\alpha(\mu x) \sum_{q=0}^{n} \frac{a_q(x)}{\mu^{2q}} + \frac{x^2}{2(\alpha+1)} j_{\alpha+1}(\mu x) \sum_{q=0}^{n-1} \frac{b_q(x)}{\mu^{2q}} + \frac{1}{x^{\alpha+1/2}}\,\theta_{\mu,n}(x),$$

where

$$j_\nu(x) = 2^\nu \Gamma(\nu+1) x^{-\nu} J_\nu(x) \quad \text{with } j_\nu(0) = 1.$$

This expansion can be deduced from Theorem 4 by using the recurrence relation [11, p. 295] for the Bessel functions in terms of Lommel polynomials. In this way we obtain

$$a_0(x)=1,$$

$$a_q(x)=\Gamma(\alpha+q+1)\left(-\frac{x^2}{4}\right)^{-q}\sum_{r=q+1}^{2q}\binom{q-1}{r-q-1}\frac{(-\frac{1}{2})^r B_r(x)}{\Gamma(\alpha+r-q+1)},\qquad q=1,2,\cdots,$$

$$b_q(x)=\frac{1}{2}\Gamma(\alpha+q+1)\left(-\frac{x^2}{2}\right)^{-q-1}\sum_{r=q+1}^{2q+1}\binom{q}{r-q-1}\frac{(-\frac{1}{2})^r B_r(x)}{\Gamma(\alpha+r-q)},\qquad q=0,1,\cdots,$$

the value of $B_0(x)$ being $2^\alpha\Gamma(\alpha+1)(C(0))^{1/2}$. These relations show how the two expansions are linked, and that they are equivalent.

Last, we give an expression of $A_2(x)$ in the general case in terms of $\chi(x)$. With $A_0(x)=1$ we have

$$A_1(x)=-\frac{1}{2}\int_0^x \chi(t)\,dt.$$

From $A_2(x)=x^2B_2(x)$ and (2.4), we derive

$$A_2(x)=\frac{1}{4}\int_0^x\left[\frac{1}{t^2}\int_0^t(s^2\chi''(s)+(1-2\alpha)s\chi'(s))\,ds+\chi(t)\int_0^t\chi(s)\,ds\right]dt$$

$$=\frac{1}{4}\int_0^x\left[\chi'(t)+(2\alpha+1)\left(\frac{1}{t^2}\int_0^t\chi(s)\,ds-\frac{1}{t}\chi(t)\right)+\chi(t)\int_0^t\chi(s)\,ds\right]dt.$$

We deduce that

$$A_2(x)=\tfrac{1}{4}\{\chi(x)-\chi(0)+2(2\alpha+1)(B_1(x)-B_1(0))+2(A_1(x))^2\}. \tag{6.5}$$

More generally, we can use the following formula for computing $A_k(x)$

$$-2B_{p+1}=H_{p+1}((\mathscr{L}+\chi)B_p)$$

where H_pf is defined by (3.3) and $\mathscr{L}$ is the differential operator defined by

$$\mathscr{L}u=u''+\frac{1-2\alpha}{x}u'.$$

Noting that $\mathscr{L}(H_pf)=H_{p+2}(\mathscr{L}f)$, we may verify by induction that

$$B_{p+1}(x)=\sum_{q=0}^{p}\left(-\frac{1}{2}\right)^{q+1}H_{p+1}(\cdots(H_{p+q+1}(\mathscr{L}^q(\chi(x)B_{p-q}(x))))\cdots).$$

6.3. Expansion of Jacobi polynomials. When $\alpha>-1$, $\beta>-1$, it is known that

$$u(x)=\frac{n!\Gamma(\alpha+1)}{\Gamma(n+\alpha+1)}P_n^{(\alpha,\beta)}(\cos x),$$

where $P_n^{(\alpha,\beta)}$ denotes the Jacobi polynomial of order n, is the solution on $(0,\pi)$ of the system

$$\frac{1}{x^{2\alpha+1}C(x)}[x^{2\alpha+1}C(x)u'(x)]'+(N^2+q(x))u(x)=0,$$

$$u(0)=1,\qquad u'(0)=0.$$

Here

$$C(x)=\left(\frac{1}{x}\sin\frac{x}{2}\right)^{2\alpha+1}\left(\cos\frac{x}{2}\right)^{2\beta+1},\qquad N=n+\frac{1}{2}(\alpha+\beta+1),$$

and

$$q(x) = -\tfrac{1}{4}(\alpha+\beta+1)^2.$$

Also, in this case $\chi(x)$ is given by

$$\chi(x) = \left(\frac{1}{4}-\alpha^2\right)\left(\frac{1}{4\sin^2(x/2)}-\frac{1}{x^2}\right)+\left(\frac{1}{4}-\beta^2\right)\frac{1}{4\cos^2(x/2)}.$$

Theorem 4 leads to the following result. If $\alpha > -\frac{1}{2}$ and $\beta > -1$, then $P_n^{(\alpha,\beta)}(\cos\theta)$ has the following expansion:

$$\left(\sin\frac{x}{2}\right)^{\alpha+1/2}\left(\cos\frac{x}{2}\right)^{\beta+1/2}P_n^{(\alpha,\beta)}(\cos x) = \frac{\Gamma(n+\alpha+1)}{n!}\left(\frac{x}{2}\right)^{1/2}\sum_{p=0}^{m}A_p(x)\frac{J_{\alpha+p}(Nx)}{N^{\alpha+p}}+x^{m+1}\mathcal{R}_{m,N}(x), \tag{6.6}$$

with

$$\mathcal{R}_{m,N} = O\left(\frac{1}{N^{\alpha+m+3/2}}\right),$$

uniformly on $[0, \pi-\varepsilon]$, ε being arbitrary. Here $A_p(x) = x^p B_p(x)$, the functions $B_p(x)$ being analytic on $[0, \pi)$ and defined recursively by (2.3). Direct computation gives

$$A_0(x) = 1,$$

$$4A_1(x) = \left(\alpha^2-\frac{1}{4}\right)\left(\frac{2}{x}-\operatorname{cotg}\frac{x}{2}\right)+\left(\beta^2-\frac{1}{4}\right)\operatorname{tg}\frac{x}{2},$$

$$4A_2(x) = \left(\alpha^2-\frac{1}{4}\right)\left(\frac{1}{x^2}-\frac{1}{4\sin^2(x/2)}+\frac{1}{12}\right)+\frac{1}{4}\left(\frac{1}{4}-\beta^2\right)\left(\frac{1}{\cos^2(x/2)}-1\right)$$
$$+\frac{1}{2}(1+2\alpha)\left(\alpha^2-\frac{1}{4}\right)\left(\frac{2}{x^2}-\frac{1}{x}\operatorname{cotg}\frac{x}{2}-\frac{1}{6}\right)+\left(\beta^2-\frac{1}{4}\right)\left(\frac{1}{x}\operatorname{tg}\frac{x}{2}-\frac{1}{2}\right)$$
$$+\frac{1}{8}\left[\left(\alpha^2-\frac{1}{4}\right)\left(\frac{2}{x}-\operatorname{cotg}\frac{x}{2}\right)+\left(\beta^2-\frac{1}{4}\right)\operatorname{tg}\frac{x}{2}\right]^2.$$

Remark. The expansion (6.6) has been obtained by Frenzen and Wong [3] in the case $\alpha > -\frac{1}{2}$, $\alpha-\beta > 2m$, $\alpha+\beta \geqq -1$, by using an integral representation due to Gasper [4]. We note that our approach is entirely different. Furthermore, by Theorem 2 we derive the following result.

THEOREM 5. *If $\alpha > -\frac{1}{2}$ and $\beta > -1$, then*

$$\left(\sin\frac{x}{2}\right)^{\alpha+1/2}\left(\cos\frac{x}{2}\right)^{\beta+1/2}P_n^{(\alpha,\beta)}(\cos x) = \frac{\Gamma(N+\alpha+1)}{n!}\left(\frac{x}{2}\right)^{1/2}\sum_{p=0}^{\infty}x^pB_p(x)\frac{J_{\alpha+p}(Nx)}{N^{\alpha+p}},$$

the infinite series being uniformly convergent on $[0, \gamma)$. Here

$$\gamma = \min\left(\frac{\pi}{2}-\varepsilon, \sqrt{\frac{2\pi-4\varepsilon}{\delta}}\right), \qquad \delta = \left(\frac{\pi}{2}-\varepsilon+\frac{2}{\pi-2\varepsilon}\right)C,$$

$$C = \max\{1+|2\alpha-1|, 2\sup\{|\chi(z)|, |z| \leqq \pi-2\varepsilon\}\},$$

ε being an arbitrary positive number.

6.4. Expansion of Jacobi functions. It is known [5] that for the Jacobi operator on $(0,\infty)$

$$(Lu)(x)=(\Delta_{\alpha,\beta}(x))^{-1}\frac{d}{dx}\left[\Delta_{\alpha,\beta}(x)\frac{du}{dx}(x)\right]+(\alpha+\beta+1)^2u(x),$$

where

$$\Delta_{\alpha,\beta}(x)=(e^x-e^{-x})^{2\alpha+1}(e^x+e^{-x})^{2\beta+1},\qquad \alpha>-\tfrac{1}{2}.$$

The system

$$Lu+\mu^2u=0,\quad \mu\in\mathbb{R},\quad u(0)=1,\quad u'(0)=0$$

has a unique solution given by

$$\varphi_\mu^{\alpha,\beta}(x)=F(\tfrac{1}{2}(\alpha+\beta+1+i\mu),\tfrac{1}{2}(\alpha+\beta+1-i\mu);\alpha+1;-\mathrm{sh}^2x),$$

where F is the hypergeometric function. In this case χ is given by

$$\chi(x)=\left(\alpha^2-\frac{1}{4}\right)\left(\frac{1}{x^2}-\frac{1}{\mathrm{sh}^2x}\right)+\left(\beta^2-\frac{1}{4}\right)\frac{1}{\mathrm{ch}^2x}.$$

THEOREM 6. *If* $\alpha>-\frac{1}{2}$, *then*

$$(\mathrm{sh}\,x)^{\alpha+1/2}(\mathrm{ch}\,x)^{\beta+1/2}\varphi_\mu^{\alpha,\beta}(x)=2^{2\alpha+\beta+1}\sum_{p=0}^{m}A_p(x)\frac{\sqrt{x}\,J_{\alpha+p}(\mu x)}{\mu^{\alpha+p}}+x^{m+1}O\left(\frac{1}{\mu^{\alpha+m+3/2}}\right),$$

where $A_p(x)=x^pB_p(x)$, *the functions* $B_p(x)$ *being analytic and defined by* (2.4).

For computing $A_p(x)$ we may use the same method as the previous case. For example, we find that

$$2A_1(x)=\left(\alpha^2-\frac{1}{4}\right)\left(\frac{1}{x}+\coth x\right)+\left(\beta^2-\frac{1}{4}\right)\mathrm{th}\,x.$$

7. Summary. The aim of this work is to show that the eigenfunction of a singular second-order differential equation can be expanded in a Bessel-function series, the coefficients being functions defined by an explicit recursive relation. This leads, without using a Paley-Wiener theorem, to an integral representation of Mehler type for the eigenfunction, with useful information on the kernel.

Furthermore, using the methods of Olver, we give an asymptotic expansion for the eigenfunction, with an upper bound for the remainder. In applications we recover asymptotic expansions for some special functions without starting from an integral representation.

Acknowledgments. The authors thank Professors K. Trimèche and A. Achour for many helpful suggestions and interesting discussions.

REFERENCES

[1] A. ACHOUR, *Opérateurs de translation et g-fonction de Littlewood-Paley associés à des opérateurs de Sturm-Liouville singuliers*, Thèse d'Etat, Faculté des Sciences de Tunis, Tunis, 1983.

[2] S. BOCHNER, *Sturm-Liouville and heat equation whose eigenfunctions are ultraspherical polynomials or associated Bessel functions*, Proceedings of the Conference on Differential Equations, University of Maryland, College Park, MD, 1956, pp. 23-48.

[3] C. L. FRENZEN AND R. WONG, *A uniform expansion of the Jacobi polynomials with error bounds*, Canad. J. Math., 37 (1985), pp. 979-1007.

[4] G. GASPER, *Formulas of Dirichlet-Mehler Type*, Lecture Notes in Math. 457, Springer-Verlag, Berlin, 1975, pp. 207-215.

[5] T. KOORNWINDER, *A new proof of a Paley-Wiener type theorem for the Jacobi transform*, Ark. Mat., 13 (1975), pp. 145-159.

[6] N. M. LEBEDEV, *Special Functions and Their Applications*, Dover, New York, 1972.

[7] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[8] R. J. STANTON AND P. A. TOMAS, *Expansion for spherical functions on noncompact symmetric spaces*, Acta. Math., 140 (1978), pp. 251-276.

[9] K. TRIMÈCHE, *Transformation intégrale de Weyl et théorème de Paley-Wiener associés à un opérateur différentiel singulier sur* $(0, \infty)$, J. Math. Pures Appl. (9), 60 (1981), pp. 51-98.

[10] ———, *Transformations intégrales de Riemann-Liouville*, Thèse d'Etat, Faculté des Sciences de Tunis, Tunis, 1981.

[11] G. N. WATSON, *A Treatise of the Theory of Bessel Functions*, Second edition, Cambridge University Press, London, New York, 1966.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1633-1641, November 1990
© 1990 Society for Industrial and Applied Mathematics
014

MONOTONIC AND OSCILLATORY SOLUTIONS OF A LINEAR NEUTRAL DELAY EQUATION WITH INFINITE LAG*

YANG KUANG† AND ALAN FELDSTEIN†

Abstract. This paper is devoted to the discussion of monotonic and oscillatory solutions of the linear neutral delay equation

$$y'(t) = Ay(t) + \sum_{i=1}^{M} B_i y(\lambda_i t) + \sum_{i=1}^{N} C_i y'(\eta_i t),$$

where $0 < \lambda_i < 1$ for $i = 1, \cdots, M$, and $0 < \eta_i < 1$ for $i = 1, \cdots, N$. Under one set of conditions, all nontrivial solutions are absolutely monotone. Under a different set of conditions, all nontrivial solutions oscillate unboundedly. This resolves most parts of the conjecture recently made by Feldstein and Jackiewicz. Some existence, uniqueness, and analyticity results are also included.

Key words. monotonic solutions, oscillatory solutions, unbounded oscillations, neutral delay equation, infinite lag, Phragmén-Lindelöf principle

AMS(MOS) subject classifications. 34K05, 34K15

1. Introduction. In a recent paper, Feldstein and Jackiewicz [6] investigate the neutral functional differential equation (where t is complex)

$$y'(t) = Ay(t) + By(\lambda t) + Cy'(\eta t) \tag{1.1}$$

where A, B, C, λ, and η are complex parameters and $0 \leqq |\lambda| < 1$, $0 \leqq |\eta| < 1$. The derivative $y'(\eta t)$ means $y'(x)$ evaluated at $x = \eta t$. Obviously, as $t \to +\infty$, the lag in (1.1) becomes infinite. At the end of their manuscript, the following conjecture was proposed.

CONJECTURE. In (1.1), suppose that $A = 0$, that B, C, λ, and η are all real, and that $-1 < C < 1$.

(a) If $B < 0$, then every nontrivial solution to (1.1) oscillates (unboundedly).

(b) If $B > 0$, then every nontrivial solution to (1.1) is monotonic.

This paper is motivated mainly by their conjecture. The equation discussed here is the following generalized version of (1.1):

$$y'(t) = Ay(t) + \sum_{i=1}^{M} B_i y(\lambda_i t) + \sum_{i=1}^{N} C_i y'(\eta_i t) \quad \text{for } t \in R. \tag{1.2}$$

The analysis presented here indicates that most parts of the above conjecture are indeed true and even true for (1.2).

When $C = 0$, equation (1.1) arises as a mathematical idealization and simplification of an industrial problem involving wave motion in the overhead supply line to an electrified railway system (see Fox et al. [8]). In this special case, (1.1) has been considered by Feldstein and Grafton [5], by Kato and McLeod [13], and by Morris, Feldstein, and Bowen [15], as well as second-order variations by Waltman [17] and Bélair [1].

As indicated in Feldstein and Jackiewicz [6], the existence of monotonic or oscillatory solutions to (1.2), while of considerable interest in its own right, is of

* Received by the editors June 19, 1989; accepted for publication (in revised form) November 29, 1989.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85287-1804. The research of the first author was partially supported by a Faculty Grant-in-Aid Award and a CLAS Summer Research Award at Arizona State University, Tempe, Arizona.

particular importance in numerical analysis because of its applicability to the development of stiffly stable numerical methods for neutral equations. See, for example, Dahlquist [4], Gear [9], and Bellen, Jackiewicz, and Zennaro [2].

This paper is organized as follows. Section 2 establishes theorems on existence, uniqueness, and analyticity of solutions to (1.2), followed by a section devoted to a discussion of conditions under which solutions to (1.2) are monotonic. Section 4 contains oscillatory and unboundedness results; Theorem 4.1 is the main result of this investigation. This paper concludes with a brief discussion and a list of some open questions.

2. Existence, uniqueness, and differentiability of solutions. This paper is devoted primarily to the discussion of monotonic and oscillatory solutions of the following linear neutral functional differential equation:

$$\begin{aligned} &y(0)=y_0,\\ &y'(t)=Ay(t)+\sum_{i=1}^{M}B_i y(\lambda_i t)+\sum_{i=1}^{N}C_i y'(\eta_i t) \quad \text{for } t\in R, \end{aligned} \tag{2.1}$$

where $0<\lambda_i<1$, $0<\eta_i<1$, and the coefficients A, B_i, and C_i are all real constants. First, this section establishes some basic results about the existence, uniqueness, and differentiability of solutions to (2.1).

For $a<b$, let $C[a, b]$ denote the complete metric space consisting of continuous functions on $[a, b]$, with the metric function ρ defined as

$$\rho(y_1(t), y_2(t))=\max_{a\leqq t\leqq b}|y_1(t)-y_2(t)|, \tag{2.2}$$

where each $y_i(t)$ is continuous on $[a, b]$. The following theorem presents conditions for the local existence and uniqueness of solutions to (2.1).

THEOREM 2.1. *Suppose that* $\alpha=\sum_{i=1}^{N}|C_i\eta_i^{-1}|<1$ *and that* $0<T<(1-\alpha)(|A|+\sum_{i=1}^{M}|B_i|)^{-1}$. *Then* (2.1) *has a unique solution on* $[0, T]$.

Proof. Existence. Since $\alpha=\sum_{i=1}^{N}|C_i\eta_i^{-1}|<1$, it follows that $\sum_{i=1}^{N}|C_i|\leqq\alpha<1$. Let

$$z_0=\left(A+\sum_{i=1}^{M}B_i\right)\left(1-\sum_{i=1}^{N}C_i\right)^{-1}y_0, \tag{2.3}$$

$$S=\{z(t)\,|\,z(0)=z_0,\ z(t)\in C[0, T]\}, \tag{2.4}$$

where $0<T<(1-\alpha)(|A|+\sum_{i=1}^{M}|B_i|)^{-1}$. Obviously, S with the metric defined in (2.2) constitutes a complete metric space.

Consider the mapping $L: S\to S$ defined as

$$Lz(t)=z_0+A\int_0^t z(s)\,ds+\sum_{i=1}^{M}B_i\int_0^{\lambda_i t}z(s)\,ds+\sum_{i=1}^{N}C_i(z(\eta_i t)-z_0). \tag{2.5}$$

Let $z_1(t)\in S$ and $z_2(t)\in S$. Denote

$$\rho(z_1, z_2)=\rho(z_1(t), z_2(t))=\max_{0\leqq t\leqq T}|z_1(t)-z_2(t)|.$$

It is easy to see that

$$\begin{aligned} Lz_1(t)-Lz_2(t)=&A\int_0^t(z_1(s)-z_2(s))\,ds+\sum_{i=1}^{M}B_i\int_0^{\lambda_i t}(z_1(s)-z_2(s))\,ds\\ &+\sum_{i=1}^{N}C_i(z_1(\eta_i t)-z_2(\eta_i t)). \end{aligned} \tag{2.6}$$

Hence, for $t \in [0, T]$,

$$|Lz_1(t) - Lz_2(t)| \leqq |A| t\rho(z_1, z_2) + \left(\sum_{i=1}^{M} \lambda_i |B_i|\right) t\rho(z_1, z_2) + \left(\sum_{i=1}^{N} |C_i|\right) \rho(z_1, z_2).$$

This implies that

$$\rho(Lz_1, Lz_2) \leqq \left[T\left(|A| + \sum_{i=1}^{M} \lambda_i |B_i|\right) + \sum_{i=1}^{N} |C_i| \right] \rho(z_1, z_2). \tag{2.7}$$

Let $\beta = T(|A| + \sum_{i=1}^{M} \lambda_i |B_i|) + \sum_{i=1}^{N} |C_i|$. Then the second hypothesis implies that

$$\begin{aligned} \beta &\leqq T\left(|A| + \sum_{i=1}^{M} |B_i|\right) + \sum_{i=1}^{N} |C_i| < 1 - \alpha + \sum_{i=1}^{N} |C_i| \\ &\leqq 1 - \sum_{i=1}^{N} |C_i| + \sum_{i=1}^{N} |C_i| = 1, \end{aligned}$$

i.e., $\beta < 1$. Therefore, L is a contraction mapping. Hence, there exists a unique $z(t) \in S$ (Waltman [18, p. 170]) such that $Lz(t) = z(t)$. This is equivalent to

$$z(t) = z_0 + A \int_0^t z(s)\, ds + \sum_{i=1}^{M} B_i \int_0^{\lambda_i t} z(s)\, ds + \sum_{i=1}^{N} C_i (z(\eta_i t) - z_0). \tag{2.8}$$

Now, let $y(t) = y_0 + \int_0^t z(s)\, ds$. It is easy to see that $y(0) = y_0$, and

$$y'(t) = Ay(t) + \sum_{i=1}^{M} B_i y(\lambda_i t) + \sum_{i=1}^{N} C_i y'(\eta_i t).$$

In other words, $y(t) = y_0 + \int_0^t z(s)\, ds$ is a solution of (2.1) on $[0, T]$.

Uniqueness. Assume $y(t)$ is a solution of (2.1) on some interval $[0, T]$, where $0 < T < (1-\alpha)(|A| + \sum_{i=1}^{M} |B_i|)^{-1}$. Then $y(t)$ must be a solution of

$$y(t) = y_0 + A \int_0^t y(s)\, ds + \sum_{i=1}^{M} B_i \lambda_i^{-1} \int_0^{\lambda_i t} y(s)\, ds + \sum_{i=1}^{N} C_i \eta_i^{-1} (y(\eta_i t) - y_0), \tag{2.9}$$

which is obtained by integrating both sides of (2.1).

Let

$$\bar{S} = \{y(t) \mid y(0) = y_0,\ y(t) \in C[0, T]\}. \tag{2.10}$$

$\bar{S}$, with the metric defined in (2.2), constitutes a complete metric space. Consider the mapping $\bar{L}: \bar{S} \to \bar{S}$ defined by

$$\begin{aligned} \bar{L}y(t) = y_0 &+ A \int_0^t y(s)\, ds + \sum_{i=1}^{M} B_i \lambda_i^{-1} \int_0^{\lambda_i t} y(s)\, ds \\ &+ \sum_{i=1}^{N} C_i \eta_i^{-1} (y(\eta_i t) - y_0). \end{aligned} \tag{2.11}$$

By an argument similar to that in the proof for the existence part, it is easy to see that $\bar{L}$ is a contraction mapping. Hence, the solution of the integral equation (2.9) is unique. This implies that there is a unique solution $y(t)$ of (2.1) and so completes the proof. □

THEOREM 2.2. *Under the same assumptions as in Theorem* 2.1, *the solution of* (2.1) *is analytic.*

Proof. For any given positive integer l, consider the differential equation

$$X(0) = y_l,$$

$$X'(t) = AX(t) + \sum_{i=1}^{M} B_i \lambda_i^l X(\lambda_i t) + \sum_{i=1}^{N} C_i \eta_i^l X'(\eta_i t), \tag{2.12}$$

where

$$y_{j+1} = \frac{A + \sum_{i=1}^{M} b_i \lambda_i^j}{1 - \sum_{i=1}^{N} C_i \eta_i^{j+1}} y_j \quad \text{for } j = 0, 1, \cdots, l-1. \tag{2.13}$$

It is easy to see that the conditions in Theorem 2.1 are satisfied for (2.12). Hence, the existence and uniqueness of its solution is guaranteed. Let $X(t)$ be such a solution. Then it can be shown by induction that $y(t)$ given by

$$y(t) = \int_0^t \left(\cdots \int_0^{s_2} \left(\int_0^{s_1} X(s)\, ds \right) ds_1 \cdots \right) ds_{l-1} + \sum_{j=0}^{l} \frac{y_j}{j!} t^j \tag{2.14}$$

is a solution of (2.1). Theorem 2.1 implies that $y(t)$ is the unique solution of (2.1). Since this is true for all l, then $y(t)$ is infinitely many times differentiable.

Consider the possible power series expression of the unique solution $y(t)$. Assume $y(t) = \sum_{l=0}^{\infty} a_l t^l$. Then

$$(l+1)a_{l+1} = \left(A + \sum_{i=1}^{M} B_i \lambda_i^l \right) a_l + \left(\sum_{i=1}^{N} C_i \eta_i^{l+1} \right) (l+1) a_{l+1}. \tag{2.15}$$

Hence

$$a_{l+1} = \frac{A + \sum_{i=1}^{M} B_i \lambda_i^l}{1 - \sum_{i=1}^{N} C_i \eta_i^{l+1}} \cdot \frac{a_l}{l+1}, \tag{2.16}$$

which leads to

$$a_l = \frac{a_0}{l!} \prod_{j=0}^{l-1} \left(\frac{A + \sum_{i=1}^{M} B_i \lambda_i^j}{1 - \sum_{i=1}^{N} C_i \eta_i^{j+1}} \right). \tag{2.17}$$

By the ratio test, it is easy to see that the series $\sum_{l=0}^{\infty} a_l t^l$ converges everywhere. Indeed, $y(t) = \sum_{l=0}^{\infty} a_l t^l$ is the unique solution of (2.1) provided that $a_0 = y_0$. Obviously, $\sum_{l=0}^{\infty} a_l t^l$ represents an analytic function. This proves the theorem. □

Remark 2.1. Since (2.1) is autonomous, it is easy to see that the local existence result in Theorem 2.1 is indeed a global one in the sense that the solution, in fact, exists for all $t \geqq 0$. This can be shown rigorously by a simple induction argument.

3. Monotonicity results. This section presents conditions under which solutions of (2.1) are monotone, or nonoscillatory. By virtue of the proof of Theorem 2.2, the following lemma is true.

LEMMA 3.1. *Suppose* $\sum_{i=1}^{N} |C_i \eta_i^{-1}| < 1$; *then the unique solution of* (2.1) *is given by*

$$y(t) = y_0 \sum_{l=0}^{\infty} \left(\prod_{j=0}^{l-1} \left(\frac{A + \sum_{i=1}^{M} B_i \lambda_i^j}{1 - \sum_{i=1}^{N} C_i \eta_i^{j+1}} \right) \right) \frac{t^l}{l!}, \tag{3.1}$$

where empty products are equal to 1.

DEFINITION 3.1. *Let* $y(t)=\sum_{l=0}^{\infty} a_l t^l$. If $d^k y(t)/dt^k > 0$ (<0) for all nonnegative integers k and for all $t>0$, then $y(t)$ is said to be *strictly absolutely monotone increasing* (*decreasing*), or simply *absolutely monotone.* (Compare with Widder [19, p. 144], who does not require the strict inequality for the derivatives.) If $d^k y(t)/dt^k \geqq 0$ $(\leqq 0)$ for all nonnegative integers k and for all $t \geqq T > 0$, then $y(t)$ is said to be *eventually absolutely monotone.*

The following lemma is obviously true.

LEMMA 3.2. *If* $y(t)=\sum_{l=0}^{\infty} a_l t^l$ *is absolutely monotone or eventually absolutely monotone, then it is nonoscillatory.*

LEMMA 3.3. *If* $y_0 \neq 0$, $A \geqq -\sum_{i=1}^{M} \min\{0, B_i\}$, *and* $\sum_{i=1}^{N} |C_i|\eta_i < 1$, *then the power series in* (3.1) *is absolutely monotone.*

Proof. Let $a_l = \prod_{j=0}^{l-1} ((A+\sum_{i=1}^{M} B_i\lambda_i^j)/(1-\sum_{i=1}^{N} C_i\eta_i^{j+1}))$. Then it is easy to see that the assumptions made in this lemma ensure that $a_l > 0$ for integer $l \geqq 0$. Note that $y(t) = y_0 \sum_{l=0}^{\infty} a_l t^l/l!$, and that $d^k y(t)/dt^k = y_0 \sum_{l=0}^{\infty} (a_{l+k} t^l)/l!$. Obviously, $d^k y(t)/dt^k > 0$ (<0) for all $k \geqq 0$ and for all $t>0$, if $y_0 > 0$ (<0). This completes the proof. □

LEMMA 3.4. *If* $A+\sum_{i=1}^{M} B_i\lambda_i^j \neq 0$ *for all integers* $j \geqq 0$, $\sum_{i=1}^{N} |C_i|\eta_i < 1$, *and* $A>0$, *then the power series in* (3.1) *is eventually absolutely monotone.*

Proof. Let $\alpha_j = (A+\sum_{i=1}^{M} B_i\lambda_i^j)/(1-\sum_{i=1}^{N} C_i\eta_i^{j+1})$ for integers $j \geqq 0$. The second hypothesis together with $0<\eta_i<1$ implies that $1-\sum_{i=1}^{N} C_i\eta_i^{j+1} > 0$. Since $0<\lambda_i<1$, there exists a nonnegative integer J, such that $A+\sum_{i=1}^{M} B_i\lambda_i^j < 0$ for $j<J$ and such that $A+\sum_{i=1}^{M} B_i\lambda_i^j > 0$ for $j \geqq J$. Without loss of generality, assume $y_0 = 1$. Then

$$y(t) = \sum_{l=0}^{J-1} \left(\prod_{j=0}^{l-1} \alpha_j\right) \frac{t^l}{l!} + \sum_{l=J}^{\infty} \left(\prod_{j=0}^{l-1} \alpha_j\right) \frac{t^l}{l!}. \tag{3.2}$$

Obviously, the second summation term on the right-hand side of (3.2) is absolutely monotone and the first summation term is a polynomial of degree $=J-1$. Therefore, there exists a sufficiently large $T>0$ such that the dominant terms in $d^k y(t)/dt^k$ will be given by $d^k/dt^k(\sum_{l=J}^{\infty} (\prod_{j=0}^{l-1} \alpha_j) t^l/l!)$ for $k \geqq 0$ and $t \geqq T$. This implies that $y(t)$ is eventually absolutely monotone. □

The combination of the above lemmas results in the following theorem.

THEOREM 3.1. *Assume that* $\sum_{i=1}^{N} |C_i\eta_i^{-1}| < 1$ *and that* $y_0 \neq 0$.

(i) *If* $A > -\sum_{i=1}^{M} \min\{0, B_i\}$, *then the unique solution of* (2.1) *is absolutely monotone.*

(ii) *If* $A>0$ *and* $A+\sum_{i=1}^{M} B_i\lambda_i^j \neq 0$, *for* $j \geqq 0$, *then the unique solution of* (2.1) *is eventually absolutely monotone.*

Proof. The proof follows immediately from the theorems in § 2 and the lemmas in this section. □

Remark 3.1. Suppose that $y_0 \neq 0$, $\varepsilon > 0$, $A>0$, $A+\sum_{i=1}^{M} B_i\lambda_i^j \neq 0$, and $\sum_{i=1}^{N} C_i\eta_i^{j+1} \neq 1$. From the power series expression of $y(t)$ in (3.1), it is then easy to see that $\lim_{t\to\infty} y(t)\, e^{-(A+\varepsilon)t} = 0$ and that $\lim_{t\to\infty} |y(t)\, e^{-(A-\varepsilon)t}| = \infty$. This indicates that $y(t)$ is eventually absolutely monotone and grows like the function e^{At}.

4. Oscillatory and unboundedness results. This section presents conditions under which solutions of (2.1) are oscillatory, or unbounded, or both.

DEFINITION 4.1. Let $f(t)$ be defined on $(-\infty, \infty)$, the *order p* of $f(t)$ is defined as

$$p = \inf\{\omega : f(t) = 0(\exp(t^{\omega})),\ |t| \to \infty\}. \tag{4.1}$$

See Titchmarsh [16, p. 248] for more details.

LEMMA 4.1. *If* $A=0$ *and* $\sum_{i=1}^{N} C_i\eta_i^{j+1} \neq 1$ *for all integers* $j \geqq 0$, *then the power series* $y(t)$ *defined by* (3.1) *has order zero.*

Proof. Let $B=\sum_{i=1}^{M}|B_i|$, $\lambda=\max\{\lambda_i \mid i=1,2,\cdots,M\}$. Then

$$|y(t)|\leqq|y_0|\sum_{k=0}^{\infty}\left(\prod_{j=0}^{k-1}\frac{B\lambda^j}{1-\sum_{i=1}^{N}C_i\eta_i^{j+1}}\right)\frac{t^k}{k!}$$

$$=|y_0|\sum_{k=0}^{\infty}\left(B^k\lambda^{(k/2)(k-1)}\left(\prod_{j=0}^{k-1}\left(1-\sum_{i=1}^{N}C_i\eta_i^{j+1}\right)\right)^{-1}\right)\frac{t^k}{k!}.$$

Obviously,

$$\lim_{j\to\infty}\sum_{i=1}^{N}C_i\eta_i^{j+1}=0.$$

Hence, for any $1>\varepsilon>0$, there exists a $K\geqq 0$, such that

$$(1-\varepsilon)^k<\left|\prod_{j=0}^{k-1}\left(1-\sum_{i=1}^{N}C_i\eta_i^{j+1}\right)\right|<(1+\varepsilon)^k \quad \text{for } k\geqq K. \tag{4.2}$$

It is known (see Titchmarsh [16, p. 253]) that the power series $\sum_{k=0}^{\infty}a_kt^k$ has a finite order p if and only if

$$\liminf_{k\to\infty}\frac{\ln(1/|a_k|)}{k\ln k}=\frac{1}{p}. \tag{4.3}$$

In the case of $y(t)$ from (3.1), $k\geqq K$ implies that

$$\lim_{k\to\infty}\frac{\ln(1/|a_k|)}{k\ln k}\geqq\lim_{k\to\infty}\frac{k(\ln(1-\varepsilon)-\ln B)-(k/2)(k-1)\ln\lambda-\ln(y_0)}{k\ln k}$$

$$=\lim_{k\to\infty}\frac{-\frac{1}{2}(k-1)\ln\lambda}{\ln k}=\infty;$$

hence, $p=0$. □

The following so-called Phragmén-Lindelöf principle can be found in § 8.73 of Titchmarsh [1958, p. 274].

PHRAGMÉN-LINDELÖF PRINCIPLE. *Let $f(z)$ be a complex function given by $f(z)=\sum_{n=0}^{\infty}\alpha_nz^n$. Let $m(r)$ denote the minimum of $|f(z)|$ on the circle $|z|=r$. If $f(z)$ has order less than $\frac{1}{2}$, then there is a sequence of values of r tending to infinity through which $m(r)\to\infty$.*

Applying this principle to the solution of (2.1) yields the lemma below. It generalizes Corollary 2 of Feldstein and Jackiewicz [6] and Theorem 5 of Morris, Feldstein, and Bowen [15].

LEMMA 4.2. *If $A=0$, $\sum_{i=1}^{N}C_i\eta_i^{i+1}\neq 1$, and $\sum_{i=1}^{M}B_i\lambda_i^j\neq 0$ for $j\geqq 0$, then the solution of* (2.1), *and all of its derivatives, is unbounded.*

Proof. Let $y(t)=\sum_{l=0}^{\infty}a_lt^l$ be the solution of (2.1). It follows from Lemma 4.1 that the order of $y(t)$ is zero. Apply the Phragmén-Lindelöf principle to $y(z)$, here z is a complex variable. Then $y(z)$ is unbounded on any ray. In particular, $y(z)$ is unbounded on the real line; that is, $y(t)$ is unbounded for the real variable t. Let

$$y_i(t)=\frac{d^iy(t)}{dt^i}=\sum_{l=0}^{\infty}\frac{(i+l)!}{i!}a_{i+l}t^l=\sum_{l=0}^{\infty}a_l^it^l$$

where

$$a_l^i=\frac{(i+l)!}{i!}a_{i+l}.$$

The same argument as in the proof of Lemma 4.1 yields

$$\lim_{k\to\infty}\frac{\ln(1/|a_k^i|)}{k\ln k}\geqq\lim_{k\to\infty}\frac{-(\ln|a_{i+k}|)-i\ln(i+k)}{k\ln k}=\lim_{k\to\infty}\frac{-\ln|a_{i+k}|}{k\ln k}$$

$$=\lim_{k\to\infty}\frac{-\frac{1}{2}(k+i)(k+i-1)\ln\lambda}{k\ln k}=\infty.$$

Hence the order of $y_i(t)$ is zero for $i=1,2,\cdots$. Thus, the Phragmén-Lindelöf principle implies that the $y_i(t)$ are all unbounded for $i=1,2,\cdots$. This proves the lemma. □

THEOREM 4.1. *In* (2.1), *assume that* $\sum_{i=1}^{N}\eta_i^{-1}\max\{0,C_i\}<1$, *that* $A=0$, *and that* $B_i<0$ *for* $i=1,2,\cdots,M$. *Then every nontrivial solution of* (2.1) *oscillates unboundedly.*

Proof. Assume $y(t)\neq 0$ is a solution of (2.1) which is not oscillatory. By the linearity of (2.1), $-y(t)$ is also a nonoscillatory solution. Therefore, without loss of generality, assume that $y(t)$ is eventually positive. That is, assume that there exists a $\bar{t}>0$ such that $y(t)>0$ for $t\geqq\bar{t}$. Let

$$r=\min_{i,j}(\lambda_i,\eta_j), \tag{4.4}$$

$$t_0=\bar{t}/r. \tag{4.5}$$

If $t\geqq t_0$, then $y(\lambda_i t)$ for $i=1,2,\cdots,M$ and $y(\eta_j t)$ for $j=1,2,\cdots,N$ are all positive. Integrating both sides of (2.1) results in

$$y(t)-\sum_{i=1}^{N}C_i\eta_i^{-1}y(\eta_i t)=y(t_0)-\sum_{i=1}^{N}C_i\eta_i^{-1}y(\eta_i t_0)+\sum_{i=1}^{M}B_i\lambda_i^{-1}\int_{\lambda_i t_0}^{\lambda_i t}y(s)\,ds. \tag{4.6}$$

Since $B_i<0$ and $y(s)>0$ for $s\in[\lambda_i t_0,\lambda_i t]$ and for $i=1,2,\cdots,M$, it follows that

$$y(t)-\sum_{i=1}^{N}C_i\eta_i^{-1}y(\eta_i t)<y(t_0)-\sum_{i=1}^{N}C_i\eta_i^{-1}y(\eta_i t_0)\quad\text{for all } t\geqq t_0. \tag{4.7}$$

Denote

$$g(t)=\min_{s}\{s\,|\,s\in[t_0,t]\text{ and such that }y(s)=\max\{y(\tau)\,|\,\tau\in[t_0,t]\}\}.$$

Clearly, $g(t)$ and $y(g(t))$ are nondecreasing, and, by Lemma 4.2,

$$\lim_{t\to\infty}g(t)=\infty\quad\text{and}\quad\lim_{t\to\infty}y(g(t))=\infty. \tag{4.8}$$

Choose t^* large enough so that $rg(t^*)>t_0$, and so that

$$y(g(t^*))>\left\{y(t_0)+\sum_{i=1}^{M}|C_i\eta_i^{-1}|y(\eta_i t_0)\right\}\left(1-\sum_{i=1}^{N}\eta_i^{-1}\max\{0,C_i\}\right)^{-1}.$$

Then for this t^*,

$$\begin{aligned}y(g(t^*))-\sum_{i=1}^{N}C_i\eta_i^{-1}y(\eta_i g(t^*))&\geqq y(g(t^*))-\sum_{i=1}^{N}\max\{0,C_i\}\eta_i^{-1}y(g(t^*))\\&=y(g(t^*))\left(1-\sum_{i=1}^{N}\eta_i^{-1}\max\{0,C_i\}\right)\\&>y(t_0)+\sum_{i=1}^{N}|C_i\eta_i^{-1}|y(\eta_i t_0).\end{aligned} \tag{4.9}$$

Obviously, (4.9) contradicts (4.7); this contradiction implies that $y(t)$ must be oscillatory. By Lemma 4.2, $y(t)$ in fact oscillates unboundedly as $t \to +\infty$. □

Theorem 4.1 clearly implies the following corollary.

COROLLARY 4.1. *If the coefficients in* (2.1) *satisfy* $A = 0$, $B_i < 0$, *and* $C_i < 0$, *then every nontrivial solution oscillates unboundedly.*

Remark 4.1. Note here that $|C_i|$ need not necessarily be less than 1. This indicates that Theorem 4.1 exceeds the expectation of Feldstein and Jackiewicz [6] in their conjecture. The proof of Theorem 4.1 for the case $N = M = 1$, $A = 0$, and $B < 0$ shows that a smaller upper bound for C may be required in order to make nontrivial solutions of (2.1) oscillate. In Theorem 4.1, this bound is η, which is smaller than the one suggested in the conjecture in Feldstein and Jackiewicz [6]. This is also in agreement with the assumptions in Theorems 2.1 and 2.2.

5. Discussion. This paper responds to the conjecture recently proposed by Feldstein and Jackiewicz [6] based on their numerical experiments. It turns out that most parts of their conjecture are indeed true, and in fact, much more has been proved in the last two sections. These results are relevant to the work of Morris, Feldstein, and Bowen [15], and those in Fox et al. [8] and Kato and McLeod [13].

When $A \neq 0$, the conclusion to Lemma 4.2 may no longer be true. That is why Theorem 4.1 assumes $A = 0$, although the proof of its conclusion seems very promising when $A < 0$. This certainly raises an interesting question to be answered. The other problems remaining to be investigated are:

(1) Can an existence and uniqueness theorem similar to Theorem 2.1 be established in the case $\sum_{i=1}^{N} |C_i \eta_i^{-1}| \geqq 1$? If this can be resolved affirmatively, then the power series solution in (3.1) is the unique solution of (2.1). In that case, the assumption $\sum_{i=1}^{N} |C_i \eta_i^{-1}| < 1$ could be deleted from Theorem 3.1.

(2) If $A < 0$ and $\sum_{i=1}^{M} B_i > 0$, then what are the conditions for which every nontrivial solution of (2.1) oscillates?

(3) What can be said if the delay functions $\lambda_i t$ and $\eta_j t$ are replaced by $\lambda_i(t - \tau_i)$ and by $\eta_j(t - \sigma_j)$, where $\tau_i \geqq 0$ and $\sigma_j \geqq 0$? When $\lambda_i = \eta_j = 1$, various results have been recently obtained by Grammatikopoulos, Grove, Ladas, and Meimaridou [10], [11] and Freedman and Kuang [7]. The method developed in Cooke and Grossman [3] may contribute to the discussion of this problem.

(4) It may be interesting to consider the case where the coefficients in (2.1) are matrix rather than scalar quantities. Aspects of such problems have been considered by Waltman [17] and Bélair [1] for the case $C_j = 0$, $M = 1$, and the order of the matrix is 2.

(5) Nonautonomous and nonlinear versions of (2.1) can also be investigated.

REFERENCES

[1] J. BÉLAIR, *Sur une équation différentielle fonctionnelle analytique*, Canad. Math. Bull., 24 (1981), pp. 43-46.

[2] A. BELLEN, Z. JACKIEWICZ, AND M. ZENNARO, *Stability analysis of one-step methods for neutral delay-differential equations*, Numer. Math., 52 (1988), pp. 605-619.

[3] K. L. COOKE AND Z. GROSSMAN, *Discrete delay, distributed delay and stability switches*, J. Math. Anal. Appl., 86 (1982), pp. 592-627.

[4] G. DAHLQUIST, *A special stability problem for linear multistep methods*, BIT, 3 (1963), pp. 27-43.

[5] A. FELDSTEIN AND C. K. GRAFTON, *Experimental mathematics: an application to retarded ordinary differential equations with infinite lag*, in Proc. 1968 ACM National Conference, Brandon Systems Press, 1968, pp. 67-71.

[6] A. FELDSTEIN AND Z. JACKIEWICZ, *Unstable neutral functional differential equations*, Canad. Math. Bull., to appear.
[7] H. I. FREEDMAN AND Y. KUANG, *Stability switches in linear scalar neutral delay equations*, Funkcial. Ekvac., to appear.
[8] L. FOX, D. F. MAYERS, J. R. OCKENDON, AND A. B. TAYLER, *On a functional differential equation*, J. Inst. Math. Appl., 8 (1971), pp. 271–307.
[9] C. W. GEAR, *The automatic integration of stiff ordinary differential equations*, Information Processing, 68, A. J. H. Morrel, ed., North-Holland, Amsterdam, 1969, pp. 187–193.
[10] M. K. GRAMMATIKOPOULOS, E. A. GROVE, AND G. LADAS, *Oscillations of first order neutral delay differential equations*, J. Math. Anal. Appl., 120 (1986), pp. 510–520.
[11] E. A. GROVE, G. LADAS, AND A. MEIMARIDOU, *A necessary and sufficient condition for the oscillation of neutral equations*, J. Math. Anal. Appl., 126 (1987), pp. 341–354.
[12] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, Berlin, 1977.
[13] T. KATO AND J. B. MCLEOD, *The function differential equation* $y'(x) = ay(\lambda x) + by(x)$, Bull. Amer. Math. Soc., 77 (1971), pp. 891–937.
[14] Y. KUANG AND A. FELDSTEIN, *Boundedness of solutions of a nonlinear nonautonomous neutral delay equation*, J. Math. Anal. Appl., to appear.
[15] G. R. MORRIS, A. FELDSTEIN, AND E. W. BOWEN, *The Phragmén-Lindelöf principle and a class of functional-differential equations*, in Ordinary Differential Equations, L. Weiss, ed., Academic Press, New York, 1972, pp. 513–540.
[16] E. C. TITCHMARSH, *The Theory of Functions*, Second ed., Oxford University Press, London, 1958.
[17] P. WALTMAN, *A note on an oscillation criterion for an equation with a retarded argument*, Canad. Math. Bull., 24 (1968), pp. 593–595.
[18] ———, *A Second Course in Elementary Differential Equations*, Academic Press, New York, 1986.
[19] D. V. WIDDER, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1946.

SIAM J. MATH. ANAL.
Vol. 21, No. 6, pp. 1642-1652, November 1990
© 1990 Society for Industrial and Applied Mathematics
015

ON MONODROMY GROUPS OF SECOND-ORDER FUCHSIAN EQUATIONS*

A. BAIDER† AND R. C. CHURCHILL†

Abstract. The paper combines classical reducibility criteria for monodromy groups of second-order Fuchsian equations with recent methods of Ziglin for establishing the nonintegrability of complex analytic Hamiltonian systems. The preliminaries on reducibility are isolated from the nonintegrability results.

Key words. monodromy group, Fuchsian equation, reducible group, hypergeometric equation, Riemann equation, Heun's equation, Hamiltonian systems, nonintegrability

AMS(MOS) subject classifications. 33A30, 33A70, 34A20, 34A30, 58F05

Introduction. The most general second-order Fuchsian equation on $\mathbf{P}^1$ with three singular points is Riemann's equation, or, in an equivalent sense, the hypergeometric equation. The monodromy groups were computed a century ago (see [5] for recent work). The most general second-order Fuchsian equation on $\mathbf{P}^1$ with four singular points is Heun's equation: only in the past decade did methods become available for computing the monodromy (see [12] and references therein), and, as far as we are aware, the results have yet to be tabulated. In fact, these groups have been explicitly calculated for very few equations (e.g., see [16]).

This is in spite of their increasing importance. For example, it is now known that when the monodromy group of an nth-order Fuchsian equation contains a solvable subgroup of finite index the equation must be "solvable by quadratures" (see [1, p. 128] or [15, p. 1096], and [9] in the case of second-order equations). A different example kindled our interest: if the monodromy group of a certain linearized equation admits no nontrivial rational integral, then an associated complex Hamiltonian system admits no meromorphic integral independent of the Hamiltonian (see § 4).

This note is concerned with extracting such information without actually computing the groups. In particular, we provide criteria which guarantee the nonintegrability of certain classes of complex Hamiltonian systems when the relevant linearized equation is Fuchsian on $\mathbf{P}^1$ with an arbitrary number of singular points. Previous techniques have been limited to the hypergeometric equation, or have required the presence of elliptic functions (e.g., see [6]). We introduce a general class of examples amenable to our methods, and illustrate the ideas with a specific problem involving Heun's equation (for related material, see [2]).

Our nonintegrability criteria require, among other things, that the relevant linearized equation be irreducible. This is a classical topic of independent interest, and for this reason the preliminaries on reducibility (§§ 1 and 2) are isolated from the material on nonintegrability (§§ 3 and 4).

1. Pseudo-Abelian subgroups of Gl (2, C). The commutator $uvu^{-1}v^{-1}$ of elements $u, v \in \mathrm{Gl}(2, \mathbf{C})$ will be denoted (u, v), and the trace and determinant of u by $\mathrm{tr}(u)$ and $\det(u)$, respectively. Pu is the Möbius transformation induced by u on $\mathbf{P}^1$.

An element $u \in \mathrm{Gl}(2, \mathbf{C})$ is *parabolic* (*generic*) if Pu has exactly one (two) fixed point(s). An equivalent condition for parabolicity is: $u \neq \lambda I$ but $\mathrm{tr}^2(u) = 4 \det(u)$. Equivalent conditions for genericity are: u has distinct eigenvalues; $\mathrm{tr}^2(u) \neq 4 \det(u)$.

* Received by the editors March 8, 1989; accepted for publication (in revised form) August 3, 1989. This research was partially supported by National Science Foundation grant DMS-8802911.

† Department of Mathematical Sciences, Hunter College, 695 Park Avenue, New York, New York 10021.

We say that elements $u, v \in \mathrm{Gl}(2, \mathbf{C})$ *quasi commute* if and only if $\mathrm{tr}((u, v)) = 2$.

PROPOSITION 1.1. *Let* $u, v \in \mathrm{Gl}(2, \mathbf{C})$.

(a) *u and v quasi commute if and only if Pu and Pv have (at least) one common fixed point, if and only if u and v can be simultaneously triangularized.*

(b) *Suppose $u \neq \lambda I \neq v$, $\lambda \in \mathbf{C}$. Then u and v commute if and only if u and v quasi commute and the fixed-point sets of Pu and Pv are identical. In particular, a generic and a parabolic element cannot commute.*

(c) *If u and v quasi commute but do not commute, then (u, v) is parabolic.*

Proof. It suffices to assume that $u, v \in \mathrm{Sl}(2, \mathbf{C})$, in which case the result is standard (e.g., see [3, Thm. 4.3.5, p. 68]). □

Let G be a subgroup of $\mathrm{Gl}(2, \mathbf{C})$ and let $G' := (G, G)$ denote the commutator subgroup. G is *pseudo-Abelian* if and only if $\mathrm{tr}\,|G' \equiv 2$. Note that:

(i) If G is pseudo-Abelian, then so is the closure of G in $\mathrm{Gl}(2, \mathbf{C})$;

(ii) Every pseudo-Abelian group is contained in a maximal one;

(iii) Every maximal pseudo-Abelian group is closed; and that

(iv) The subgroup T of $\mathrm{Gl}(2, \mathbf{C})$ of lower triangular matrices is maximal pseudo-Abelian. Indeed, by Proposition 1.1(a) any element quasi commuting with $\left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$ must have $\left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$ as an eigenvector.

PROPOSITION 1.2. *The following statements concerning a subgroup G of $\mathrm{Gl}(2, \mathbf{C})$ are equivalent*:

(a) *G is pseudo-Abelian.*

(b) *G is reducible, i.e., conjugate to a subgroup of the triangular group T (see* (iv) *above).*

Proof. Only (a)⇒(b) requires proof, so assume G is pseudo-Abelian. We consider two cases separately: (i) G contains a parabolic element; (ii) G admits no such element.

(i) Let $g \in G$ be parabolic. By Proposition 1.1(a) any $x \in G$ has the property that Px fixes the (unique) fixed point of Pg. An eigenvector of g is therefore an eigenvector of every element of G. But then G is triangular in any basis involving that common eigenvector.

(ii) In this case G is Abelian; otherwise some (u, v) must be parabolic, which is a contradiction. Unless G is diagonal, in which case the result is obvious, G must contain a generic element u, and since G is Abelian every other element of G must preserve the eigendirections of u. But then G is diagonal with respect to a basis of eigenvectors of u. □

COROLLARY 1.3. *The maximal pseudo-Abelian subgroups of $\mathrm{Gl}(2, \mathbf{C})$ are precisely the conjugates of T.*

COROLLARY 1.4. *Pseudo-Abelian subgroups of $\mathrm{Gl}(2, \mathbf{C})$ are solvable.*

Proof. The commutator subgroup of T is Abelian. □

COROLLARY 1.5. *Elements $u, v \in \mathrm{Gl}(2, \mathbf{C})$ quasi commute if and only if they generate a pseudo-Abelian subgroup.*

A simple modification of the proof of Proposition 1.2 leads to the following standard result.

PROPOSITION 1.6. *A subgroup of $\mathrm{Gl}(2, \mathbf{C})$ is commutative if and only if it is conjugate to either*

(a) *A diagonal group, or*

(b) *A subgroup of*

$$\left\{ \begin{bmatrix} \lambda & 0 \\ a & \lambda \end{bmatrix} : \lambda \neq 0 \right\}.$$

Everything stated thus far holds with $\mathrm{Gl}(2, \mathbf{C})$ replaced by $\mathrm{Sl}(2, \mathbf{C})$. For the remainder of the section we restrict our attention to the latter group.

With each subset T of $\mathbf{C}$ let $\mathbf{Q}(T)$ denote the subfield of $\mathbf{C}$ generated by T. Now associate with each subset S of $\mathrm{Sl}(2,\mathbf{C})$ the following objects:

(a) $G(S)$, the subgroup generated by S;

(b) $\mathbf{Q}(\mathrm{tr}(S))$, where $\mathrm{tr}(S) := \{\mathrm{tr}(u): u \in S\}$; and

(c) $\mathbf{Q}(\sigma(S))$, where $\sigma(S) := \bigcup \{\sigma(u)$ (the spectrum of u): $u \in S\}$.

For $g \in S$ and $\lambda \in \sigma(g)$ the relation $\lambda^2 - \mathrm{tr}(g)\lambda + 1 = 0$ implies that $\mathbf{Q}(\mathrm{tr}(S)) \subseteq \mathbf{Q}(\sigma(S))$, and that the extension is algebraic. We thus have a diagram

$$\begin{array}{ccc} \mathbf{Q}(\sigma(S)) & \to & \mathbf{Q}(\sigma(G(S))) \\ \uparrow & & \uparrow \\ \mathbf{Q}(\mathrm{tr}(S)) & \to & \mathbf{Q}(\mathrm{tr}(G(S))) \end{array}$$

of field extensions, with vertical arrows algebraic.

THEOREM 1.7. *If $G(S) \subseteq \mathrm{Sl}(2,\mathbf{C})$ is a pseudo-Abelian group, then $\mathbf{Q}(\sigma(S)) = \mathbf{Q}(\sigma(G(S)))$. In particular, $\mathbf{Q}(\sigma(G(S)))$ is algebraic over $\mathbf{Q}(\mathrm{tr}(S))$; hence the trace of any element of $G = G(S)$ is algebraic over the field generated by the traces of the generators of G.*

Proof. By Proposition 1.2 we may assume S is contained in the triangular group T, in which case the result is obvious. □

COROLLARY 1.8. *Suppose $s_1, \cdots, s_{n+1} \in \mathrm{Sl}(2,\mathbf{C})$ satisfy*

(a) $\prod_{j=1}^{n+1} s_j = I$, *and*

(b) $\mathrm{tr}(s_{n+1})$ *is transcendental over* $\mathbf{Q}(\mathrm{tr}(\{s_1, \cdots, s_n\}))$.

Then $G := G(\{s_1, \cdots, s_n\})$ is not pseudo-Abelian, and hence not Abelian.

2. Pseudo-Abelian monodromy groups. Let X be a Zariski open set of projective space $\mathbf{P}^1$ with complement $\{\alpha_1, \cdots, \alpha_m, \alpha_{m+1} = \alpha_\infty = \infty\}$, and let ν_j denote the usual order function [7, p. 130] on the space of germs of meromorphic functions at α_j. The most general nth order Fuchsian equation on X has the form

$$y^{(n)} + c_1(x)y^{(n-1)} + \cdots + c_n(x)y = 0, \tag{2.1}$$

where the rational functions c_j satisfy

$$\begin{aligned} &\nu_j(c_k) \geqq -k, \qquad 1 \leqq j \leqq m, \quad 1 \leqq k \leqq n, \quad \text{and} \\ &\nu_\infty(c_k) \geqq k, \qquad 1 \leqq k \leqq n. \end{aligned} \tag{2.2}$$

In the terminology of algebraic geometry the relations (2.2) are equivalent to $c_k \in L(kD)$, where D is the divisor $\alpha_1 + \alpha_2 + \cdots + \alpha_m - \alpha_\infty$ (cf., e.g., [7, p. 130]). Each element c_k of $L(kD)$ can be uniquely written in the form

$$c_k(x) = \frac{p(x)}{((x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_m))^k}, \tag{2.3}$$

where $p(x)$ is a polynomial of degree at most $k(m-1)$. Since the map $c_k \to p$ is an isomorphism between $L(kD)$ and the space of all such polynomials, $\dim(L(kD)) = k(m-1)+1$. We will regard the finite-dimensional vector space $L_n = \prod_{k=1}^{n} L(kD)$ as the parameter space for all Fuchsian equations on X of the form (2.1), and will denote the elements by $\vec{\mathbf{c}} = (c_1, \cdots, c_n)$, where $c_k \in L(kD)$.

Define $\lambda_j (= \lambda_j^k) \in L(kD)^*$ by

$$\lambda_j(c_k) := \lim_{x \to \alpha_j} (x - \alpha_j)^k c(x), \qquad j = 1, \cdots, m,$$

$$\lambda_\infty(c_k) := \lim_{x \to \infty} x^k c_k(x).$$

Remark 2.4. The meaning of the λ_j at the finite poles can be understood from the partial fraction expansion

$$c_k = \sum_{j=1}^{m} \frac{\lambda_j}{(x-\alpha_j)^k} + \cdots,$$

where the dots represent lower-order poles at the α_j, and $\lambda_j = \lambda_j(c_k)$. Writing c_k in the form (2.3) and assuming the identification $c_k \approx p$, we immediately see that for finite α_j the linear form λ_j is a constant multiple of the evaluation map (Dirac delta) at α_j, and λ_∞ is the leading coefficient of p. This shows that the λ_j for $j = 1, \cdots, m$ are linearly independent. In the special case when $k = 1$ or $m = 1$, we have $\dim L(kD) = m$, so that $\{\lambda_j\}_{j=1}^m$ is a basis of $L(kD)$, and direct calculation yields $\lambda_\infty = \sum_{j=1}^m \lambda_j$. However, *in all other cases* λ_∞ *is independent of the remaining* $\{\lambda_j\}_{j=1}^m$.

Henceforth we assume $m \geqq 2$, i.e., that the number of singular points is at least 3, and to study monodromy we fix a point $x_0 \in X$. Analytic continuation of solution germs of (2.1) along loops based at x_0 yields a right representation of the homotopy group $\pi_1(X, x_0)$ on the n-dimensional vector space V of solution germs of (2.1) at x_0. In discussing this representation we will not distinguish loops and their corresponding homotopy classes. For $\gamma \in \pi_1(X, x_0)$ let $s(\gamma; \vec{c}) \in \mathrm{Gl}(V, \mathbf{C})$ be the corresponding monodromy element. In view of analytic dependence on parameters, the function $\vec{c} \to s(\gamma; \vec{c})$ from L_n to $\mathrm{Gl}(V, \mathbf{C})$ is entire. For each j choose a positively oriented loop γ_j enclosing α_j and no other singularity. As is well known, $\pi_1(X, x_0)$ is freely generated by the γ_j corresponding to finite singularities, and we may assume (perhaps after a permutation) that the relation $\gamma_1 \cdots \gamma_m \gamma_\infty = 1$ holds. This implies the analogous relation

$$s_1(\vec{c}) \cdots s_m(\vec{c}) s_\infty(\vec{c}) = 1 \tag{2.5}$$

at the monodromy level, where $s_j(\vec{c}) := s(\gamma_j; \vec{c})$.

Now we restrict ourselves to second-order linear equations, and write (2.1) as

$$L(y) := y'' + c_1(x)y' + c_2(x)y = 0. \tag{2.6}$$

THEOREM 2.7. *Let G be the monodromy group of* (2.6). *Then the following statements are equivalent*:

(a) *G is pseudo-Abelian*;

(b) *G is reducible*;

(c) $L = L_1 \circ L_2$, *where the operators* L_j *correspond to first-order Fuchsian equations.*

The reducibility of a Fuchsian equation can always be effectively determined [13, §§ 176–178]; special cases are considered in Theorem 2.24 and its corollaries. The result here goes back to Beke [4] (see also [8]).

Proof. The equivalence of (a) and (b) is a restatement of Proposition 1.2. Now for any local solution z of $L(z) = 0$, we readily check that L admits the (local) factorization $L = L_1 \circ L_2$, where $L_1 := d/dx + r + c_1$, $L_2 := d/dx - r$, and $r := z'/z$. Assuming (b), we choose z to be a common eigenvector for G. In that case r is G-invariant and therefore single-valued on the punctured sphere. Moreover, the Fuchsian character of L implies that the singularities of r are at worst (simple) poles. Thus L_1 and L_2 are Fuchsian, and the above factorization is global.

Conversely, if $L = L_1 \circ L_2$, then G has a one-dimensional invariant subspace, namely, the space of solutions of $L_2(y) = 0$. □

With reference to (2.6), set $a_j := \lambda_j^1(c_1)$, $b_j := \lambda_j^2(c_2)$, $j = 1, \cdots, m+1$. Then the characteristic exponents at a_j are

$$\begin{aligned} &\tfrac{1}{2}(1 - a_j \pm \sqrt{(a_j-1)^2 - 4b_j}), \qquad j = 1, \cdots, m, \\ &\tfrac{1}{2}(1 - a_\infty \pm \sqrt{(a_\infty - 1)^2 - 4b_\infty}), \end{aligned} \tag{2.8}$$

from which we compute the traces of the monodromy elements corresponding to the generating loops γ_j to be

$$(2.9)\qquad \begin{aligned} t_j &= -2\, e^{-\pi i a_j} \cos \pi\sqrt{(a_j-1)^2-4b_j}, \qquad j=1,\cdots,m,\\ t_\infty &= -2\, e^{-\pi i a_\infty} \cos \pi\sqrt{(a_\infty-1)^2-4b_j}. \end{aligned}$$

Remark 2.10. Since the function $\cos \pi\sqrt{z}$ is entire, the traces (2.9) are entire functions on the parameter space L_2. This is in agreement with our earlier remark concerning the analyticity of $\vec{\mathbf{c}} \to s(\gamma; \vec{\mathbf{c}})$.

The "normal form" of (2.6) is

$$(2.11)\qquad y'' + c(x)y = 0, \qquad c(x) := c_2 - (c_1/2)^2 - (c_1)'/2,$$

and is obtained from (2.6) after multiplying the dependent function by a conveniently chosen (generally multivalued) function of x. Thus the projective representations of the monodromy of both equations are the same. More to the point here is the fact that the monodromy groups of equations (2.6) and (2.11) are simultaneously Abelian or non-Abelian, and the same holds for the pseudo-Abelian property. The advantage of (2.11) is that the monodromy is a subgroup of $\mathrm{Sl}(2, \mathbf{C})$, allowing application of the results of § 1. Identities (2.9) now become

$$(2.12)\qquad \begin{aligned} t_j &= -2\cos\left(\pi\sqrt{(a_j-1)^2-4b_j}\right), \qquad j=1,\cdots,m,\\ t_\infty &= -2\cos\left(\pi\sqrt{(a_\infty-1)^2-4b_\infty}\right), \end{aligned}$$

where the a_j, b_j are computed in terms of the coefficients of (2.6).

Remark 2.13. The last set of trace equations exhibit a remarkable property which becomes apparent when contrasted with the relation (2.5) satisfied by the corresponding monodromy matrices. While $s_\infty(\vec{\mathbf{c}})$ is a function of the $s_j(\vec{\mathbf{c}})$ associated with finite singularities, t_∞ *is independent of the remaining* t_j. Indeed, it follows from the italicized statement at the end of Remark 2.4 that while $a_\infty = \sum_{j=1}^{m} a_j$, b_∞ is linearly independent of those b_j with $1 \leqq j \leqq m$.

Examples 2.14.

(a) The *hypergeometric equation.* In (2.6) let

$$c_1(x) = \frac{\gamma - (\alpha+\beta+1)x}{x(1-x)} \quad \text{and} \quad c_2(x) = -\frac{\alpha\beta}{x(1-x)},$$

where $\alpha, \beta, \gamma \in \mathbf{C}$ are arbitrary. Using $\alpha_1 = 0$, $\alpha_2 = 1$, $\alpha_3 = \infty$, (2.12) gives

$$t_1 = -2\cos \pi(\gamma-1), \quad t_2 = -2\cos \pi(\alpha+\beta-\gamma), \quad t_3 = -2\cos \pi(\alpha-\beta),$$

for (2.11).

(b) *Riemann's equation.* In (2.6) let

$$c_1(x) = \frac{(1-\eta_1-\mu_1)}{x} + \frac{(1-\eta_2-\mu_2)}{x-1}, \quad \text{and}$$

$$c_2(x) = \frac{\eta_1\mu_1}{x^2} + \frac{\eta_2\mu_2}{(x-1)^2} + \frac{(\eta_3\mu_3 - \eta_1\mu_1 - \eta_2\mu_2)}{x(x-1)},$$

where $\eta_j, \mu_j \in \mathbf{C}$ are arbitrary other than $\sum(\eta_j + \mu_j) = 1$. Using $\alpha_1 = 0$, $\alpha_2 = 1$, $\alpha_3 = \infty$, (2.12) gives

$$t_j = -2\cos \pi(\eta_j - \mu_j), \qquad j = 1, 2, \infty,$$

for (2.11).

(c) *Heun's equation.* In (2.6) let

$$c_1(x)=\frac{\gamma}{x}+\frac{\delta}{x-1}+\frac{\varepsilon}{x-a} \quad \text{and} \quad c_2(x)=\frac{(\alpha\beta x-1)}{x(x-1)(x-a)},$$

where α, β, γ, δ, ε, and $a\in\mathbf{C}$ are arbitrary other than $\alpha+\beta+1=\gamma+\delta+\varepsilon$, $|a|\geqq 1$, and $a\neq 1$. For $\alpha_1=0$, $\alpha_2=1$, $\alpha_3=a$, and $\alpha_4=\infty$, (2.12) gives

$$t_1=2\cos\pi\gamma, \quad t_2=2\cos\pi\delta, \quad t_3=2\cos\pi\varepsilon, \quad t=-2\cos\pi(\alpha-\beta),$$

for (2.10).

THEOREM 2.15. *Suppose t_∞ in* (2.12) *is transcendental over $\mathbf{Q}(t_1,\cdots,t_m)$. Then the monodromy of* (2.6) *is not pseudo-Abelian (hence not Abelian).*

Proof. In view of (2.5), Corollary 1.7 applies to the monodromy of the normalized equation and thus to (2.6). □

COROLLARY 2.16. *Given an arbitrary $c_1\in L(D)$ there exist $c_2\in L(2D)$ such that the monodromy of* (2.6) *is not pseudo-Abelian.*

Proof. To obtain the proof use Remark 2.13 in conjunction with Theorem 2.15. □

COROLLARY 2.17. *The monodromy of the second-order Fuchsian equation* (2.6) *is generically nonpseudo-Abelian. More precisely, there is an analytic subvariety V of the parameter space such that $\vec{\mathbf{c}}\in V$ if the monodromy of* (2.6) *is pseudo-Abelian.*

Proof. By the previous corollary there exist loops γ, γ' such that the entire function $f(\vec{\mathbf{c}}):=\operatorname{tr}(s(\gamma;\vec{\mathbf{c}}),s(\gamma';\vec{\mathbf{c}}))$ is not a constant. Set $V:=\{\vec{\mathbf{c}}: f(\vec{\mathbf{c}})=2\}$. □

COROLLARY 2.18. *The general Fuchsian equation of order $n\geqq 2$ with more than two singular points is generically non-Abelian.*

Proof. "Prolongation" (i.e., differentiation $n-2$ times) of (2.6) gives rise to a particular Fuchsian equation of type (2.1) with non-Abelian monodromy if that of (2.6) is so. Choose loops γ, γ' so that the map $\vec{\mathbf{c}}\to(s(\gamma;\vec{\mathbf{c}}),s(\gamma';\vec{\mathbf{c}}))$ is not constant on $L_2=\prod_{k=1}^{2}L(kD)$. Since the identity map injects solutions of (2.6) into solutions of (2.1), it is clear that for the same loops the corresponding map on $L_n=\prod_{k=1}^{n}L(kD)$ is also nonconstant. □

Specific illustrations of Theorem 2.15 can be constructed using Examples 2.14 and the following remark.

Remark 2.19. $\cos z$ *is transcendental over* $\mathbf{C}$ *if* $z\in\mathbf{C}$ *is algebraic.* Indeed, $e^{iz}=\cos z+i\sin z$ is transcendental by the result of Hermite-Lindemann (e.g., see [10, p. 681]).

We will give an explicit description of the conditions in Theorem 2.15 and its corollaries for equations with three and four singular points. Let

$$p_{ij}(x):=x^2-t_it_jx+t_i^2+t_j^2-4, \tag{2.20}$$

where t_j is as in (2.12).

PROPOSITION 2.21. *s_i and s_j quasi commute if and only if* $\operatorname{tr}(s_is_j)$ *is a zero of* $p_{ij}(x)$.

Proof. The result follows from the standard Fricke-Klein formula (e.g., see [11, p. 703]) $\operatorname{tr}((u,v))=\operatorname{tr}^2(uv)-\operatorname{tr}(u)\operatorname{tr}(v)\operatorname{tr}(uv)+\operatorname{tr}^2(u)+\operatorname{tr}^2(v)-2$, $u,v\in \mathrm{Sl}(2,\mathbf{C})$. □

PROPOSITION 2.22. *The monodromy group of a second-order Fuchsian equation with three singular points is pseudo-Abelian if and only if t_3 is a zero of $p_{12}(x)$ or, equivalently, if and only if*

$$t_1^2+t_2^2+t_3^2-t_1t_2t_3-4=0. \tag{2.23}$$

Proof. $s_3=(s_1s_2)^{-1}$ by (2.5), and therefore $t_3=\operatorname{tr}(s_1s_2)$. Now apply Proposition 2.21 and Corollary 1.5. □

We shall now derive an explicit condition equivalent to (2.23).

THEOREM 2.24. *A second-order Fuchsian equation with three singular points is reducible if and only if at least one of the (up to sign) four determinations of the expression*

$$((a_1-1)^2-4b_1)^{1/2}+((a_2-1)^2-4b_2)^{1/2}+((a_3-1)^2-4b_3)^{1/2} \tag{2.25}$$

is an odd integer.

Proof (following a suggestion by the referee). Choose a specific square root A_j of $(a_j-1)^2-4b_j$, $j=1,2,3$, and observe that instead of (2.25) it suffices to consider the four quantities $A_1+A_2+A_3$, $A_1+A_2-A_3$, $A_1-A_2+A_3$, and $-A_1+A_2+A_3$. Introduce quantities u_j, $j=1,2,3$ solving

$$u_1+u_2+1=A_3, \quad u_2+u_3=A_1, \quad u_3+u_1=A_2. \tag{2.26}$$

By (2.12) $t_j=-2\cos\pi A_j$, so that using (2.26) the left-hand side of (2.23) becomes $4(\cos^2\pi(u_1+u_2)+\cos^2\pi(u_2+u_3)+\cos^2\pi(u_3+u_1)-2\cos\pi(u_1+u_2)\cos\pi(u_2+u_3)\cos\pi(u_3+u_1)-1)=$ (by a trigonometric identity) $=-16(\sin(\pi u_1)\sin(\pi u_2)\cdot\sin(\pi u_3)\sin(\pi(u_1+u_2+u_3)))$. Thus (2.23) holds if and only if one of u_1, u_2, u_3, and $u_1+u_2+u_3$ is an integer. Solving (2.26) we obtain

$$u_1=\tfrac{1}{2}(A_2+A_3-A_1-1), \qquad u_2=\tfrac{1}{2}(A_3+A_1-A_2-1),$$
$$u_3=\tfrac{1}{2}(A_1+A_2-A_3+1), \qquad u_1+u_2+u_3=\tfrac{1}{2}(A_1+A_2+A_3-1),$$

from which the result readily follows. □

COROLLARY 2.27. *The hypergeometric equation is reducible if and only if at least one of α, β, $\gamma-\beta$, and $\gamma-\alpha$ is an integer.*

Proof. The expression (2.25) becomes $\pm(\gamma-1)\pm(\alpha+\beta-\gamma)\pm(\alpha-\beta)$, and at least one of these must be an odd integer. □

Corollary 2.27 was already known to Frobenius, who used Kummer's table of solutions for the hypergeometric equation to establish it. For another "algebraic" proof, see [14].

COROLLARY 2.28. *Riemann's equation is reducible if and only if at least one of $\eta_1+\eta_2+\eta_3$, $\eta_1+\eta_2+\mu_3$, $\eta_1+\mu_2+\eta_3$, and $\mu_1+\eta_2+\eta_3$ is an integer.*

Proof. The expression (2.25) becomes $\sum_{j=1}^{3}\pm(\eta_j-\mu_j)$. The result follows because of $\sum(\eta_j+\mu_j)=1$. □

With $p_{ij}(x)$ as given by (2.20) we have Proposition 2.29.

PROPOSITION 2.29. *The monodromy group of a second-order Fuchsian equation with four singular points is pseudo-Abelian only if $p_{12}(x)$ and $p_{34}(x)$ have a common zero.*

Proof. $s_3s_4=(s_1s_2)^{-1}$ by (2.5); hence $\operatorname{tr}(s_1s_2)=\operatorname{tr}(s_3s_4)$. Now apply Proposition 2.21. □

Remark 2.30. Using the resultant of the two polynomials in Proposition 2.29, we can write down a condition in terms of the traces similar to (2.23). These conditions constitute an explicit description of the variety V of Corollary 2.17.

3. Ziglin groups. Let $D\subseteq \mathrm{Sl}(2,\mathbf{C})$ consist of those elements which preserve the coordinate axes of $\mathbf{C}^2$; let P consist of those which permute these axes. Thus

$$D:=\left\{\begin{bmatrix}\lambda & 0\\ 0 & \lambda^{-1}\end{bmatrix}:\lambda\in\mathbf{C}_*\right\},$$
$$P:=\left\{\begin{bmatrix}0 & -\mu^{-1}\\ \mu & 0\end{bmatrix}:\mu\in\mathbf{C}_*\right\},$$

where $\mathbf{C}_*:=\mathbf{C}\backslash\{0\}$. $D\cup P$ is a group, and D is a normal subgroup.

PROPOSITION 3.1. *Suppose* $u, v \in D \cup P$.

(a) *If* $v \in P$ *then* $\operatorname{tr}(v) = 0$ *and* $v^2 = -I$.

(b) *If* $u \in D$ *then* $(u, v) = I$ *if* $v \in D$, *and* $(u, v) = u^2$ *if* $v \in P$.

The proof is a straightforward verification. □

If $\mathbf{e} \subseteq \mathbf{C}^2$ is a basis and $u \in \mathrm{Sl}(2, \mathbf{C})$, denote the corresponding matrix of u by $u_{\mathbf{e}}$.

THEOREM 3.2. *Let* $u, v \in \mathrm{Sl}(2, \mathbf{C})$ *with* u *generic, and suppose* $(u, v) \neq I$. *Then* v *permutes the eigenspaces of* u *if and only if* (u, v) *preserves these spaces. Moreover, if this is the case then* $(u, v) = u^2$.

Proof. Let $\mathbf{e}$ be an eigenbasis of u. If v permutes the eigenspaces then $v_{\mathbf{e}} \in P$, whereas $u_{\mathbf{e}} \in D$. But then $(u, v) = u^2$ by Proposition 3.1(b); hence (u, v) preserves the eigenspaces.

To prove the converse let $0, \infty \in \mathbf{P}^1$ be the eigendirections of u, and set $a := Pu$, $b := Pv$, $c := P(a, b)$. If $x \in \{0, \infty\}$ then $x = cx = aba^{-1}b^{-1}x$ implies $b^{-1}a^{-1}x = a^{-1}b^{-1}x$, and therefore $b^{-1}x = aa^{-1}b^{-1}x = ab^{-1}a^{-1}x = ab^{-1}x$. Thus $b^{-1}x \in \{0, \infty\}$, and so $b\{0, \infty\} = \{0, \infty\}$. Since $(u, v) \neq I$, we conclude that v permutes the eigenspaces. □

A subgroup G of $\mathrm{Sl}(2, \mathbf{C})$ is a DP-*group* if there are two distinct one-dimensional subspaces of $\mathbf{C}^2$ which are either preserved or permuted by each element of G. Equivalently, G is conjugate to a subgroup of $D \cup P$.

PROPOSITION 3.3. *Suppose* $S \subseteq \mathrm{Sl}(2, \mathbf{C})$ *is a set of generators for a* DP-*group and* $\operatorname{tr}(s) \neq 0$ *for each* $s \in S$. *Then* G *is diagonalizable, and therefore Abelian.*

Proof. The proof is obtained by Proposition 3.1. □

If $u \in \mathrm{Sl}(2, \mathbf{C})$ has spectrum $\{\lambda, \lambda^{-1}\}$, then

$$\operatorname{tr}(u^n) = \lambda^n + \lambda^{-n}, \qquad n \in \mathbf{Z}. \tag{3.4}$$

When λ, λ^{-1} are roots of unity, u is *resonant*. Nonresonant elements are obviously generic, and by (3.4) have infinite order. Since

$$\lambda = \tfrac{1}{2}(t - \sqrt{t^2 - 4}), \qquad t = \operatorname{tr}(u),$$

resonance can be determined directly from $\operatorname{tr}(u)$. Note that u is nonresonant if $\operatorname{tr}(u)$ is transcendental (see Remark 2.19).

A subgroup G of $\mathrm{Sl}(2, \mathbf{C})$ is a *Ziglin group* if there is a nonconstant rational function on $\mathbf{C}^2$ which is preserved by the action of G.

THEOREM 3.5. *Let* $G \subseteq \mathrm{Sl}(2, \mathbf{C})$ *be a group with a nonresonant element. Then* G *is a* DP-*group if and only if* G *is a Ziglin group. Moreover, the forward implication is true without the nonresonance hypothesis.*

Proof. Suppose G is a DP-group and let x, y be the usual coordinate functions on $\mathbf{C}^2$. Then $z = (xy)^2$ is preserved by the action of G, and the forward implication follows by conjugation.

The converse is a particular case of a result of Ziglin (see [17, Thm. 2, p. 182]). For the sake of completeness we give a short proof only valid in the two-dimensional case. First observe that if G has a rational integral $f = p(x, y)/q(x, y)$ (i.e., invariant function), then it also has a homogeneous integral (for example, the quotient of the lowest homogeneous components of p and q). Now a homogeneous integral is expressible as $f = \prod_{k=1}^{n} l_k^{j_k}$ where the $l_k = a_k x + b_k y$ are linear forms and the j_k are nonzero integers, so that the finite set $\{l_k = 0\}_{k=1}^{n}$ of projective space is PG-invariant, and, in particular, Pu-invariant where u is a nonresonant element of G. But the only finite nonempty invariant subsets of Pu have either one or two elements, as they consist of eigendirections of u. Thus n is at most 2, and $n = 1$ is impossible by nonresonance. Changing coordinates we can write $l_1 = x$ and $l_2 = y$. The invariance of f now implies

that for any $v \in G$, either $v^*x = \lambda x$ or $v^*x = \mu y$. In the first case $v^*y = \lambda^{-1}y$ and in the second $v^*y = -\mu^{-1}x$. □

THEOREM 3.6. *Suppose a group $G \subseteq \mathrm{Sl}(2, \mathbf{C})$ contains a nonresonant element u and an element v such that $I \neq (u, v) \neq u^2$. Then G is not a Ziglin group.*

Proof. The proof is obtained by Theorem 3.5 and Proposition 3.1. □

The importance of knowing when a subgroup $G \subseteq \mathrm{Sl}(2, \mathbf{C})$ is not a Ziglin group will be discussed in § 4 (cf. Theorem 4.5). Theorem 3.6 has been the standard method for establishing this fact (e.g., see [6] and references therein). We offer the following alternative.

THEOREM 3.7. *Suppose $s_1, \cdots, s_{m+1} \in \mathrm{Sl}(2, \mathbf{C})$ satisfy*

(a) $\prod_{j=1}^{m+1} s_j = 1$,

(b) $\mathrm{tr}(s_j) \neq 0$, $j = 1, \cdots, m$, *and*

(c) t_{m+1} *is transcendental over* $\mathbf{Q}(\{t_1, \cdots, t_m\})$.

Then $G = G(\{s_1, \cdots, s_m\})$ is not a Ziglin group.

Proof. Otherwise G is a DP-group by Theorem 3.5, and must be Abelian by (b) and Proposition 3.3. But (c) then contradicts Corollary 1.8.

There is an "eigenvalue" version of Theorem 3.7.

THEOREM 3.8. *Suppose $s_1, \cdots, s_{m+1} \in \mathrm{Sl}(2, \mathbf{C})$, and let $\lambda_j^{\pm 1}$ be the eigenvalues of s_j, $j = 1, \cdots, m+1$. Assume*

(a) $\prod_{j=1}^{m+1} s_j = 1$,

(b) *some s_j is nonresonant,*

(c) $\mathrm{tr}(s_j) \neq 0$, $j = 1, \cdots, m$, *and*

(d) $\prod_{j=1}^{n+1} \lambda_j^{\pm 1} \neq 1$.

Then $G = G(\{s_1, \cdots, s_m\})$ is not a Ziglin group.

Proof. Otherwise G is diagonalizable (Theorem 3.5 and Proposition 3.3), and in such a context we easily see that (a) and (d) are inconsistent. □

4. Nonintegrability. On $\mathbf{C}^4 \approx \{(x, x_2, y, y_2)\}$ with symplectic structure $dx \wedge dy + dx_2 \wedge dy_2$ consider a meromorphic Hamiltonian of the form

$$H(x, x_2, y, y_2) = h(x, y) + \frac{1}{2} p(x, y) x_2^2 + \frac{1}{2} \frac{\partial h}{\partial y}(x, y) y_2^2 + O_3(x_2, y_2), \tag{4.1}$$

where

$$p(x, y) = a(x) \frac{\partial c}{\partial y}(x, y), \qquad h(x, y) = b(x) c(x, y). \tag{4.2}$$

The xy-plane is tangent to the vector field X_H associated with H, and if we endow that plane with the symplectic structure $dx \wedge dy$, then the restriction is precisely the vector field X_h.

A phase curve Γ of X_H contained in the xy-plane must also be contained within the analytic set defined by

$$b(x) c(x, y) = E, \tag{4.3}$$

where E is the energy of the curve. For example, if $b(x)$ is a separable polynomial of degree $2g+1$ or $2g+2$, $1 \leqq g \in \mathbf{Z}$, $c(x, y) = y^{-2}$ and $E \neq 0$, then Γ lies on a surface of genus g. If Γ is maximal then it must be a component of (4.3) after singular points have been removed.

The tangency assumption on the xy-plane implies that the variational equation along Γ decouples into two sets of equations. One set is immediately identified as the

variational equation of Γ as a phase curve of X_h; the other is

$$\frac{d}{dt}\begin{bmatrix}\xi_2\\ \eta_2\end{bmatrix}=\begin{bmatrix}0 & \partial h/\partial y\\ -p(x,y) & 0\end{bmatrix}\begin{bmatrix}\xi_2\\ \eta_2\end{bmatrix}, \tag{4.4}$$

and is called the *normal variational equation* (NVE) *along* Γ.

THEOREM 4.5 (Ziglin [17]). *If X_H admits a meromorphic integral in a neighborhood of Γ which is independent of H, then the monodromy group of the* NVE *is a Ziglin group.*

Using (4.2), and the observation that on Γ we have

$$dt=-\frac{dy}{\partial h/\partial x}=\frac{dx}{\partial h/\partial y},$$

equation (4.4) assumes the simpler form

$$dw=\begin{bmatrix}0 & 1\\ -a(x)/b(x) & 0\end{bmatrix}dx.$$

We can therefore view the NVE as the pull-back of

$$\xi''+(a(x)/b(x))\xi=0 \qquad (' = d/dx) \tag{4.6}$$

under the mapping $\rho:(x,y)\in\Gamma\to x\in\mathbf{C}$. As a consequence the monodromy of the NVE embeds into that of (4.6), and is identical when ρ is injective (see [6, § 2]).

Example 4.7. In (4.1) let

$$p(x,y)=p(x)=\left(\frac{1}{4\pi^2}\right)(\pi^2-\eta^2)x-1, \qquad h(x,y)=x(x^2-1)y,$$

where $\eta\in\mathbf{C}$ is arbitrary. *If η is algebraic the vector field X_H associated with* (4.1) *is nonintegrable.*

Indeed, in terms of (4.2) we have $a(x)=(1/4\pi^2)(\pi^2-\eta^2)x-1$, $b(x)=x(x^2-1)$, and $c(x,y)\equiv y$. Letting $E=0$ in (4.3) we can take $\Gamma=\mathbf{P}^1\backslash\{0,1,-1,\infty\}$, in which case ρ is the identity map, and (4.6) is

$$\xi''+\frac{(1/4\pi^2)(\pi^2-\eta^2)x-1}{x(x^2-1)}x=0. \tag{4.8}$$

Defining $\alpha,\beta\in\mathbf{C}$ by

$$\alpha+\beta=-1, \qquad \alpha-\beta=\eta/\pi,$$

we have $\alpha\beta=\frac{1}{4}((\alpha+\beta)^2-(\alpha-\beta)^2)=(1/4\pi^2)(\pi^2-\eta^2)$, and (4.8) becomes

$$\xi''+\frac{\alpha\beta x-1}{x(x^2-1)}x=0, \qquad \alpha+\beta+1=0.$$

This is Heun's equation with $\gamma=\delta=\varepsilon=0$, and so (see Example 2.14(c))

$$t_1=t_2=t_3=2, \qquad t_\infty=-2\cos\pi(\alpha-\beta).$$

If $\eta=\pi(\alpha-\beta)$ is algebraic, then t_∞ is transcendental by Remark 2.19, and the monodromy group cannot be a Ziglin group by Theorem 3.7. The assertion now follows from Theorem 4.5.

When the mapping $\rho:(x,y)\in\Gamma\to x\in\mathbf{C}$ is not injective Theorem 3.7 does not apply. The situation is investigated extensively in [2].

Acknowledgments. The work was initiated while the authors were members of the Mathematical Sciences Research Institute, University of California at Berkeley. It is a pleasure to thank the director and staff there for providing such a stimulating research environment. Conversations with W. Nelson and D. L. Rod on this material are also gratefully acknowledged.

REFERENCES

[1] D. V. ANOSOV AND V. I. ARNOL'D, EDS., *Dynamical Systems* I, in The Encyclopedia of Mathematical Sciences, Vol. 1, Springer-Verlag, Berlin, New York, 1988.

[2] A. BAIDER, R. C. CHURCHILL, AND D. L. ROD, *Monodromy and non-integrability in complex Hamiltonian systems*, J. Dynamics Differential Equations, to appear.

[3] A. F. BEARDON, *The Geometry of Discrete Groups*, Springer-Verlag, Berlin, New York, 1983.

[4] E. BEKE, *Die Irreducibilität der homogene Differentialgleichungen*, Math. Ann., 45 (1894), pp. 278-294.

[5] F. BEUKERS AND G. HECKMAN, *Monodromy for the hypergeometric function* ${}_nF_{n-1}$, Invent. Math., 95 (1989), pp. 325-354.

[6] R. C. CHURCHILL AND D. L. ROD, *Geometrical aspects of Ziglin's non-integrability theorem for complex Hamiltonian systems*, J. Differential Equations, 76 (1988), pp. 91-114.

[7] P. GRIFFITHS AND J. HARRIS, *Principles of Algebraic Geometry*, John Wiley, New York, 1978.

[8] E. R. KOLCHIN, *Algebraic matric groups and the Picard-Vessiot theory of homogeneous linear ordinary differential equations*, Ann. of Math., 49 (1948), pp. 1-42.

[9] J. KOVACIC, *An algorithm for solving second order linear homogeneous equations*, J. Symbolic Comput., 2 (1986), pp. 3-43.

[10] S. LANG, *Algebra*, Second edition, Addison-Wesley, Reading, MA, 1984.

[11] W. MAGNUS, *Monodromy groups and Hill's equation*, Comm. Pure Appl. Math., 29 (1976), pp. 701-716.

[12] R. SCHAFKE AND D. SCHMIDT, *The connection problem for general linear ordinary differential equations at two regular singular points with applications to the theory of special functions*, SIAM J. Math. Anal., 11 (1980), pp. 848-862.

[13] L. SCHLESINGER, *Handbuch der Theorie der linearen Differentialgleichungen*, Teubner, Leipzig, 1897.

[14] M. SETOYANAGI, *Transcendental Liouville solutions of hypergeometric differential equations*, Kyoto Sangyo University, Kyoto, Japan, preprint.

[15] M. SINGER AND M. D. TRETKOFF, *A classification of differential equations of Fuchsian class*, Amer. J. Math., 107 (1985), pp. 1093-1109.

[16] K. TAKANO AND E. BANNAI, *A global study of Jordan-Pochammer differential equations*, Funkcial. Ekvac., 19 (1976), pp. 85-99.

[17] S. L. ZIGLIN, *Branching of solutions and non-existence of first integrals in Hamiltonian mechanics* I, Functional Anal. Appl., 16 (1982), pp. 181-189.